Improved Tabu Search and Simulated Annealing Methods For Nonlinear Data Assimilation

Elias D. Nino-Ruiz[a], Xin-She Yang[b]

[a] *Applied Mathematics and Computer Science Laboratory*
Department of Computer Science,
Universidad del Norte, Barranquilla 080001, Colombia.
[b] *School of Science and Technology,*
Middlesex University, London NW4 4BT, United Kingdom.

Abstract

Nonlinear data assimilation can be a very challenging task. Four local search methods are proposed for nonlinear data assimilation in this paper. The methods work as follows: At each iteration, the observation operator is linearized around the current solution, and a gradient approximation of the three dimensional variational (3D-Var) cost function is obtained. Then, samples along potential steepest descent directions of the 3D-Var cost function are generated, and the acceptance/rejection criteria for such samples are similar to those proposed by the Tabu Search and the Simulated Annealing framework. In addition, such samples can be drawn within certain sub-spaces so as to reduce the computational effort of computing search directions. Once a posterior mode is estimated, matrix-free ensemble Kalman filter approaches can be implemented to estimate posterior members. Furthermore, the convergence of the proposed methods is theoretically proven based on the necessary assumptions and conditions. Numerical experiments have been performed by using the Lorenz-96 model. The numerical results show that the cost function values on average can be reduced by several orders of magnitudes by using the proposed methods. Even more, the proposed methods can converge faster to posterior modes when sub-space approximations are employed to reduce the computational efforts among iterations.

Citation details: E.D. Nino-Ruiz and X.-S. Yang, Improved Tabu Search and Simulated Annealing Methods For Nonlinear Data Assimilation, *Applied Soft Computing*, Volume 83, October 2019, 105624.
https://doi.org/10.1016/j.asoc.2019.105624
(Accepted 8 July 2019, Available online 18 July 2019)

Keywords: Nonlinear Observation Operator, Data Assimilation, Tabu Search, Simulated Annealing, Ensemble Kalman Filter

URL: `https://sites.google.com/a/vt.edu/eliasnino/` (Elias D. Nino-Ruiz)

2010 MSC: 49K10, 49M05, 49M15

1. Introduction

In sequential Data Assimilation (DA), the forecasts of imperfect numerical models are calibrated to real noisy observations so as to roughly estimate the actual state of a dynamical system $\boldsymbol{\psi}^* \in \mathbb{R}^{n\times 1}$ [1] where the main physical and dynamical processes are approximately modelled by

$$\boldsymbol{\psi}^*_{\text{next}} = \mathcal{M}_{t_{\text{current}}\to t_{\text{next}}}\left(\boldsymbol{\psi}^*_{\text{current}}\right) ,$$

where n is the dimension of the model state or the model resolution. The model $\mathcal{M}(.)$ is an imperfect, numerical model, and its underlying error distributions of forecasts are approximated by normal distributions

$$\boldsymbol{\psi} \sim \mathcal{N}\left(\boldsymbol{\psi}^b, \mathbf{B}\right) , \tag{1}$$

where the background state $\boldsymbol{\psi}^b \in \mathbb{R}^{n\times 1}$ is the prior estimate of $\boldsymbol{\psi}^*$ before any observations or measurements become available, and $\mathbf{B} \in \mathbb{R}^{n\times n}$ is the covariance matrix of the background errors. In addition, the observations are also treated as random variables with Gaussian errors

$$\mathbf{o} \sim \mathcal{N}\left(\mathcal{W}\left(\boldsymbol{\psi}\right), \widehat{\mathbf{R}}\right) , \tag{2}$$

where $\mathbf{o} \in \mathbb{R}^{m\times 1}$ are the m measurements or observations, and $\widehat{\mathbf{R}} \in \mathbb{R}^{m\times m}$ is the covariance matrix of measurement errors. Here, the observation operator $\mathcal{W} : \mathbb{R}^{n\times 1} \to \mathbb{R}^{m\times 1}$ maps the observations to their corresponding model spaces. From the Bayesian rule, it can be shown [2, 3] that the state which maximizes the posterior probability can be obtained by

$$\mathcal{G}(\boldsymbol{\psi}) = \frac{1}{2}\cdot\left\|\boldsymbol{\psi}-\boldsymbol{\psi}^b\right\|^2_{\mathbf{B}^{-1}} + \frac{1}{2}\cdot\left\|\mathbf{o}-\mathcal{W}\left(\boldsymbol{\psi}\right)\right\|^2_{\widehat{\mathbf{R}}^{-1}} , \tag{3}$$

which is the three-dimensional variational (3D-Var) cost function. Thus, the state $\boldsymbol{\psi}^a \in \mathbb{R}^{n\times 1}$ to best-fit the given data can be estimated through the solution of the following 3D-Var optimization problem:

$$\boldsymbol{\psi}^a = \arg\min_{\boldsymbol{\psi}} \mathcal{G}(\boldsymbol{\psi}) . \tag{4}$$

Though this optimization problem is in general nonlinear, however, when the observation operator is linear, a closed-form expression for computing $\boldsymbol{\psi}^a$ in Eq. (4) is possible, especially in terms of the ensemble Kalman filtering (EnKF) context [4, 5, 6].

In general, Netwon-like methods can be employed for numerically solving optimization problems of the form (4) when observation operators are nonlinear. In most cases, the observation operator is linearized within some small

neighbourhood of $\boldsymbol{\psi}^b$ from which a gradient approximation of Eq. (3) can be computed so as to choose a suitable along the steepest descent direction. This iterative process is repeated until a predefined stopping criterion is met. Unfortunately, there are just a few works in this direction and even more, their theoretical convergence are missed [7, 8].

On the ohter hand, stochastic method such as sampling via the Markov Chain Monte Carlo (MCMC) can be used as well. However, such sampling methods may become inefficient under the current operational settings [9, 10], due to the so-called curse of dimensionality [11]. For the present problem, we believe the stochastic algorithms such as Local Search methods can be used so as to estimate the posterior modes of error distributions. In this context, the challenge is to find a transition function which can rapidly allow methods to reach regions of search spaces where the values of the 3D-Var cost function in Eq. (3) become small. Such transition functions can be defined by using gradient approximations to Eq. (3). Then, states can be proposed potentially along steepest descent directions of the above defined 3D-Var cost function.

The outline of this paper is as follows. Section 2 briefly introduces the concepts about data assimilation as well as local search methods, and Section 3 presents four local search methods for nonlinear data assimilation wherein transition functions are defined over descent direction approximations concerning the 3D-Var cost function. Then, Section 4 performs some numerical experiments in order to assess the accuracy of the proposed methods by using the Lorenz 96 model and different configurations for the tests. Finally, Section 5 concludes with discussions for further research.

2. Problems and Formulations

2.1. Data Assimilation: The Ensemble Kalman Filter

In order to estimate the moments of prior error distributions such as Eq. (1), an ensemble of model of realizations is used in terms of the ensemble Kalman filter (EnKF) [12]. For an ensemble size N, we have

$$\boldsymbol{\Psi}^b = \left[\boldsymbol{\psi}^{b[1]}, \boldsymbol{\psi}^{b[2]}, \ldots, \boldsymbol{\psi}^{b[N]}\right] \in \mathbb{R}^{n\times N}\,, \tag{5a}$$

so that

$$\boldsymbol{\psi}^b \approx \overline{\boldsymbol{\psi}}^b = \frac{1}{N}\cdot\sum_{e=1}^{N}\boldsymbol{\psi}^{b[e]} \in \mathbb{R}^{n\times 1}\,, \tag{5b}$$

and

$$\mathbf{B} \approx \mathbf{P}^b = \frac{1}{N}\cdot\boldsymbol{\Delta\Psi}\cdot\boldsymbol{\Delta\Psi}^T \in \mathbb{R}^{n\times n}\,, \tag{5c}$$

where $\boldsymbol{\psi}^{b[e]} \in \mathbb{R}^{n\times 1}$ is the e-th ensemble member for the e-th ensemble (with $1 \leq e \leq N$). Here, $\boldsymbol{\Delta\Psi} \in \mathbb{R}^{n\times N}$ is the matrix of perturbations, which can be

calculated by

$$\mathbf{\Delta\Psi} = \mathbf{\Psi}^b - \overline{\boldsymbol{\psi}}^b \cdot \mathbf{1}^T \,, \tag{5d}$$

where $\mathbf{1}$ is a constant unit vector of the same dimension with all its components being ones. In the EnKF, the main analysis step [13, 14] is performed by virtually solving a 3D-Var optimization problem for each prior member in (5a) [15, 2]. Thus, for a given observation $\mathbf{o} \in \mathbb{R}^{m\times 1}$, the posterior ensemble can be computed as follows [16, 6]:

$$\mathbf{\Psi}^a = \mathbf{\Psi}^b + \mathbf{P}^a \cdot \mathbf{\Delta Y} \in \mathbb{R}^{n\times N} \,, \tag{6a}$$

where $\mathbf{P}^a \in \mathbb{R}^{n\times n}$ is a low-rank approximation of the posterior covariance matrix in the following form:

$$\mathbf{P}^a = \left[\left[\mathbf{P}^b\right]^{-1} + \mathbf{W}^T \cdot \widehat{\mathbf{R}}^{-1} \cdot \mathbf{W}\right]^{-1} . \tag{6b}$$

The innovation matrix $\mathbf{\Delta Y}$ on the synthetic observations can be written as

$$\mathbf{\Delta Y} = \mathbf{W}^T \cdot \widehat{\mathbf{R}}^{-1} \cdot \left[\mathbf{o} \cdot \mathbf{1}^T + \mathbf{E} - \mathbf{W} \cdot \mathbf{\Psi}^b\right] \in \mathbb{R}^{m\times N} \,,$$

where each column of matrix $\mathbf{E} \in \mathbb{R}^{m\times N}$ follows a multivariate standard normal distribution, which makes the filter statistically consistent, but sampling noise can be induced during the assimilation step. In the operational Data Assimilation (DA), high-resolution models often requires the ensemble sizes to be hundreds (due to the computational effort involved in a single model propagation), and the sampling noise can thus degrade the quality of analysis corrections on the prior members (5a). An immediate consequence is that the ensemble covariance matrix (5c) becomes low-rank [17] and subsequently spurious correlations can impact the quality of analysis innovations in (6a). A common strategy in order to counteract this effect is to use a proper localization method. Such methods can can dissipate long-distance correlations by using one of the three techniques (in the current literature): covariance matrix localization, spatial domain localization, and observation localization.

In the covariance localization method, the structure of ensemble covariances (5c) can be achieved by componentwise multiplications with a so-called *localization matrix* whose structure typically mitigates long-distance correlations among model components (grid components in space). Another possible choice is to estimate sparse precision covariance matrices whose structure can rely on the conditional independence among different model components with regard to their physical distances. With the above conditions, the ensemble Kalman filter, based on a modified Cholesky decomposition, is proposed [5], namely, the EnKF-MC. In the EnKF-MC, the analysis step can be summarized as follows [4]:

$$\mathbf{\Psi}^a = \mathbf{\Psi}^b + \widehat{\mathbf{A}} \cdot \mathbf{\Delta Y}, \tag{7a}$$

with

$$\widehat{\mathbf{A}} = \left[\widetilde{\mathbf{B}}^{-1} + \mathbf{W}^T \cdot \widehat{\mathbf{R}}^{-1} \cdot \mathbf{W}\right]^{-1} \in \mathbb{R}^{n \times n}\,, \tag{7b}$$

which is a well-conditioned estimate of the posterior covariance matrix by means of the modified Cholesky decomposition [18]:

$$\widetilde{\mathbf{B}}^{-1} = \mathbf{L}^T \cdot \mathbf{D} \cdot \mathbf{L} \in \mathbb{R}^{n \times n}\,, \tag{7c}$$

The matrix $\mathbf{L} \in \mathbb{R}^{n \times n}$ is sparse and its structure is lower triangular with elements:

$$\{\mathbf{L}\}_{ij} = \begin{cases} -\varsigma_{ij}\,, & \text{for } j \in Z(i,\vartheta), \\ 1\,, & \text{for } i = j, \\ 0\,, & \text{otherwise}, \end{cases}$$

where the parameters ς_{ij} are obtained by fitting the following linear models:

$$\boldsymbol{\psi}^{[i]} - \sum_{j \in Z(i,\vartheta)} \boldsymbol{\psi}^{[j]} \cdot \varsigma_{ij} + \boldsymbol{\gamma}_i \in \mathbb{R}^{N \times 1} = \mathbf{0}\,.$$

Here,$\boldsymbol{\gamma}_i \in \mathbb{R}^{n \times 1}$ is normally distributed, and $\boldsymbol{\psi}^{[i]} \in \mathbb{R}^{N \times 1}$ corresponds to the i-th transposed row of the ensemble (5a). In addition, $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a diagonal matrix given by

$$\{\mathbf{D}\}_{ii} = \widehat{\mathbf{var}}\left(\boldsymbol{\psi}^{[i]} - \sum_{j \in Z(i,\vartheta)} \boldsymbol{\psi}^{[j]} \cdot \varsigma_{ij}\right)^{-1}\,, \text{ for } 1 \leq i \leq n\,,$$

where $\widehat{\mathbf{var}}$ stands for the empirical variance. $Z(i,\vartheta)$ is a set storing the predecessor indices of model component i for a given radius of influence ϑ, subject to some ordering of model components. An example in a two-dimensional domain is shown in Fig. 1.

(a) For component 6 in this model, the blue region forms its local neighborhood (a box) when $\vartheta = 1$.

(b) For the same component 6, the blue region is considered as its predecessors when $\vartheta = 1$.

Figure 1: Local grid components and local grid predecessors for component 6 in the grid when $\vartheta = 1$. Model variables are ordered by means of a column-major format.

In the current literature, methods have been proposed in order to avoid the direct computation of (7b) by exploiting the special structure of (7c). This is usually achieved via Sherman Morrison based formulas [2] and/or rank-one

updates over Cholesky factors [19, 20] without matrix inversion and computation of posterior members. Once the analysis ensemble is obtained, the analysis members are propagated until new observations are available, and then a new background ensemble is obtained. This process is repeated until all observations within the assimilation window are assimilated.

For technical details about the domain and observation localization methods, further discussions about these topics can be found in [21, 22].

In the case when the observation operator is non-linear, EnKF formulations can struggle to obtain adequate estimates of the posterior moments (and error distributions). In such cases, stochastic sampling methods becomes preferable over ensemble based methods. For instance, MCMC methods [23, 24, 25] are commonly used to sample complex probability density functions in low dimensional spaces. However, in the context of DA, the required number of samples for successfully approaching posterior moments increases exponentially [26] with respect to the number of parameters under consideration.

Some recent efforts have focused on some accelerating MCMC methods for non-Gaussian data assimilation, for instance, either by modifying proposal functions [27] or ether by using Verlet integrators [28, 29]. Thus, there is a need to carry out further research so as to apply such methods under current operational DA scenarios.

2.2. Local Search Methods

In order to cope with nonlinear obseration operation for assimilation, Local Search (LS) methods can be investigated in this DA context. Recent studies have successfully applied such LS methods for solving inverse problems [30, 31], which are a general class of DA problems. In general, most LS methods attempt to explore the search space Γ (space of feasible solutions) by using a so-called transition function $F : \Gamma \times \Upsilon \to \Gamma$, which enable to calculate the state $x' \in \Gamma$ from another $x \in \Gamma$ by

$$x' = F(x, \theta)\,.$$

Here, $\theta \in \Upsilon$ is a set of additional parameters defined in the space Υ (i.e., $\mathbb{R}$). Furthermore, x is known as the *current state*, while x' is commonly referred to as the *proposed state*. For example, the transition function for $\boldsymbol{\psi} \in \mathbb{R}^{n\times 1}$ and $\boldsymbol{\delta\psi} \in \mathbb{R}^{n\times 1}$

$$\boldsymbol{\psi}' = F(\boldsymbol{\psi}, \boldsymbol{\delta\psi}) = \boldsymbol{\psi} + \boldsymbol{\delta\psi}, \text{ for } \boldsymbol{\delta\psi} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right), \tag{8}$$

can be naively chosen to solving an optimization problem in the form of (4). In this case, both the search space and the parameter space are the same: $\mathbb{R}^{n\times 1}$. The acceptance/rejection criterion of proposed states vary from method to method. For instance, in the Tabu Search (TS) method [32, 33], new states $\boldsymbol{\psi}'$ are preferred over current ones $\boldsymbol{\psi}$ as long as [34] the following condition is met:

$$\mathcal{G}\left(\boldsymbol{\psi}'\right) \leq \mathcal{G}\left(\boldsymbol{\psi}\right)\,. \tag{9}$$

Briefly speaking, a general TS framework for solving the 3D-Var optimization problem by using the transition function (8) can be summarized as the Algorithm 1.

Algorithm 1 Tabu Search method for solving 3D-Var optimization problems.

Require: Initial solution $\boldsymbol{\psi}^{(0)}$, typically $\boldsymbol{\psi}^{(0)} \leftarrow \overline{\boldsymbol{\psi}}^{b}$.
Ensure: A posterior mode approximation $\overline{\boldsymbol{\psi}}^{a}$ of Eq. (4).
1: **for** $u = 0 \rightarrow U$ **do**
2: Draw $\boldsymbol{\delta\psi}^{(u)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
3: Compute $\mathbf{z}^{(u)} = F\left(\boldsymbol{\psi}^{(u)}, \boldsymbol{\delta\psi}^{(u)}\right)$ via Eq. (8).
4: **if** $\mathcal{G}\left(\mathbf{z}^{(u)}\right) \leq \mathcal{G}(\boldsymbol{\psi})$ **then**
5: $\boldsymbol{\psi}^{(u+1)} \leftarrow \mathbf{z}^{(u)}$
6: **else**
7: $\boldsymbol{\psi}^{(u+1)} \leftarrow \boldsymbol{\psi}^{(u)}$
8: Set $\overline{\boldsymbol{\psi}}^{a} \leftarrow \boldsymbol{\psi}^{(u)}$.

Some TS implementations make use of so-called tabu lists [32] in order to circumvent cycles during optimization steps. In the context of DA, a tabu list may not be practical, given huge search-space dimensions (i.e., vector state sizes range in the order of millions). Simulated annealing (SA) inspired approaches are another related family of well-known LS methods [35, 36, 37]. In these methods, the acceptance/rejection criterion (18) is replaced by a probabilistic one via the Boltzmann probability distribution:

$$\delta\left(\boldsymbol{\psi}', \boldsymbol{\psi}\right) = \min\left(1, \exp\left(-\left[\frac{\mathcal{G}\left(\boldsymbol{\psi}'\right) - \mathcal{G}\left(\boldsymbol{\psi}\right)}{T}\right]\right)\right), \tag{10}$$

where the T parameter is the temperature which varies as iterations. A higher value of T lead to a higher acceptance rate so that proposed states with large cost function values may be accepted as current solutions. But it may leads to slow convergence and even runs the risk of getting trapped in non-stationary points, which can be avoided through some modifications of relevant acceptance/rejection rules (i.e., by having a near one cooling factor). The process is repeated until a stopping criterion is met. For instance, a minimum temperature $T_{\min}$ can be imposed as a lower bound (an user-defined parameter). During the iterations, the temperature is updated based on a cooling schedule via a cooling factor $0 < \rho < 1$, typically $\rho \in [0.8, 0.95]$. A general framework of SA for solving the optimization problem (4) by using the transition function (8) can be summarized as the Algorithm 2.

Algorithm 2 Simulated annealing method for optimizing 3D-Var problems.

Require: Initial solution $\boldsymbol{\psi}^{(0)}$, typically $\boldsymbol{\psi}^{(0)} \leftarrow \overline{\boldsymbol{\psi}}^{b}$, initial temperature T_{ini}, cooler factor ρ, lowest temperature $T_{\min}$.
Ensure: A posterior mode approximation $\overline{\boldsymbol{\psi}}^{a}$ of Eq. (4).

$T \leftarrow T_{\text{ini}}$
$u \leftarrow 0$
while $T > T_{\min}$ **do**
  Draw $\boldsymbol{\delta\psi}^{(u)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Compute $\mathbf{z}^{(u)} = F\left(\boldsymbol{\psi}^{(u)}, \boldsymbol{\delta\psi}^{(u)}\right)$ via Eq. (8).
  Draw $\gamma \sim \mathcal{U}[0, 1]$. ▷ $\mathcal{U}$ stands for Uniform distribution.
  if $\gamma < \delta\left(\mathbf{z}^{(u)}, \boldsymbol{\psi}^{(u)}\right)$ **then**
    $\boldsymbol{\psi}^{(u+1)} \leftarrow \mathbf{z}^{(u)}$
  else
    $\boldsymbol{\psi}^{(u+1)} \leftarrow \boldsymbol{\psi}^{(u)}$
  $T \leftarrow \rho \cdot T$
  $u \leftarrow u + 1$
Set $\overline{\boldsymbol{\psi}}^{a} \leftarrow \boldsymbol{\psi}^{(u)}$.

There are many other effective LS methods proposed in the current literature [38, 39, 40], which we do not discuss here due to the limitation of space. A comprehensive survey of those methods can be found in [41, 42, 43].

2.3. Gradient Based Optimization Techniques and Convergence Properties

In nonlinear numerical optimization [44, 45], optimization problems of the form (4) are commonly solved by iterative schemes such as

$$\boldsymbol{\psi}^{(u+1)} = \boldsymbol{\psi}^{(u)} + \boldsymbol{\delta\psi}^{(u)}, \tag{11}$$

wherein $\boldsymbol{\delta\psi}^{(u)}$ is a search direction, often along the steepest descent direction [46, 47, 48, 49]

$$\boldsymbol{\delta\psi}^{(u)} = -Z_1 \cdot \nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right), \tag{12a}$$

where $Z_1 \in \mathbb{R}$ is a constant which makes the computation (11) (physically) consistent. This is usually achieved by the Newton's step [50, 51, 52]

$$\nabla^2\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right) \cdot \boldsymbol{\delta\psi}^{(u)} = -\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right), \tag{12b}$$

or a quasi-Newton based method [53, 54, 55],

$$\mathbf{P}^{(u)} \cdot \boldsymbol{\delta\psi}^{(u)} = -\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right), \tag{12c}$$

where $\mathbf{P}^{(u)} \in \mathbb{R}^{n\times n}$ is a positive definite matrix. A concise survey of Newton based methods can be found in [56].

Another relevant family of methods proposed in the current literature are the reduced-space approximations [57, 58, 59]. In this framework, search directions $\boldsymbol{\delta\psi}^{(u)}$ are constrained to the space spanned by a pre-defined set of basis vectors $\boldsymbol{\Phi}^{(u)} \in \mathbb{R}^{n\times K}$, and thus iterations commonly take the form:

$$\boldsymbol{\psi}^{(u+1)} = \boldsymbol{\psi}^{(u)} + \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}\,, \tag{12d}$$

where the weights $\boldsymbol{\mu} \in \mathbb{R}^{K\times 1}$ can be computed by solving the optimization problem

$$\boldsymbol{\mu}^* = \arg\min_{\boldsymbol{\mu}} \mathcal{G}\left(\boldsymbol{\psi}^{(u)} + \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}\right). \tag{13}$$

It is worth pointing out that step sizes in (12) can be too large, their optimal length can be approximated by using line search methods [60, 61, 62], which can ensure global convergence of iterative processes to stationary points defined by first order optimality conditions. This holds as long as some assumptions over functions, gradients, and (potentially) Hessians are preserved [63]. In such line search methods, the following assumptions are commonly used:

A The function $f(\boldsymbol{\psi})$ has a lower bound on $\Omega_0 = \{\boldsymbol{\psi} \in \mathbb{R}^{n\times 1}, f(\boldsymbol{\psi}) \leq f(\boldsymbol{\psi}_0)\}$, where $\boldsymbol{\psi}_0 \in \mathbb{R}^{n\times 1}$ is available.

B The gradient $\nabla f(\boldsymbol{\psi})$ is assumed to be Lipschitz continuous on an open convex set B, containing Ω_0,

$$\|\nabla\mathcal{G}(\boldsymbol{\psi}) - \nabla\mathcal{G}(\mathbf{z})\| \leq L\cdot\|\boldsymbol{\psi}-\mathbf{z}\|\,,\ \text{ for } \boldsymbol{\psi},\,\mathbf{z}\in B, \text{ and } L>0.$$

All the above conditions, together with the iterative form

$$\boldsymbol{\psi}^{(u+1)} = \boldsymbol{\psi}^{(u)} + \alpha\cdot\boldsymbol{\delta\psi}^{(u)}\,, \tag{14}$$

can ensure global convergence [64], as long as α is chosen approximately as a minimizer

$$\alpha^* = \arg\min_{\alpha\geq 0} \mathcal{G}\left(\boldsymbol{\psi}^{(u)} + \alpha\cdot\boldsymbol{\delta\psi}^{(u)}\right). \tag{15}$$

In principle, this optimization problem (15) can be partially solved by well-known rules for choosing step sizes in the context of line search [49].

We believe that it is advantageous to combine stochastic methods and gradient approximations of (3), which enables the solution of the 3D-Var optimization problem (4) successfully. The convergence of such methods can be proved via common assumptions in the context of gradient-based optimization methods. In the next section, we will explore some of these ideas.

3. Proposed Methods

Following the formulations in the previous section, we now propose four LS methods for solving the 3D-Var optimization problem (4). In all cases, the initial seed $\boldsymbol{\psi}^{(0)}$ of our iterative methods is the background ensemble mean $\overline{\boldsymbol{\psi}}^{b}$ (5b). Let u the u-th iteration, for $1 \leq u \leq U$, where U is the maximum number of iterations. The main rationale behind our approach is somehow to obtain at least one mode of the posterior error distribution.

3.1. Tabu Search Single Gradient Approximation

At iteration u, the Tabu Search Single Gradient Approximation (TS-SGA) in essence proceeds as follows. The observation operator is first linearized about the current solution $\boldsymbol{\psi}^{(u)}$:

$$\mathcal{W}\left(\boldsymbol{\psi}\right) \approx \mathcal{G}\left(\boldsymbol{\psi}\right) = \mathcal{W}\left(\boldsymbol{\psi}^{(u)}\right) + \mathbf{W}_{\boldsymbol{\psi}^{(u)}} \cdot \left[\boldsymbol{\psi} - \boldsymbol{\psi}^{(u)}\right] , \tag{16a}$$

where its Jacobian matrix $\mathbf{W}_{\boldsymbol{\psi}^{(u)}}$ of $\mathcal{W}(\boldsymbol{\psi})$ is given by

$$\mathbf{W}_{\boldsymbol{\psi}^{(u)}} = \frac{\partial}{\partial \boldsymbol{\psi}} \left\{\mathcal{W}(\boldsymbol{\psi})\right\} |_{\boldsymbol{\psi}=\boldsymbol{\psi}^{(u)}} \in \mathbb{R}^{m \times n} .$$

Then, the objective or cost function (3) can be approximated by a quadratic form:

$$\widehat{\mathcal{G}}(\boldsymbol{\psi}) = \frac{1}{2} \cdot \left\| \boldsymbol{\psi} - \overline{\boldsymbol{\psi}}^{b} \right\|^{2}_{\widetilde{\mathbf{B}}^{-1}} + \frac{1}{2} \cdot \left\| \mathbf{o} - \mathcal{G}\left(\boldsymbol{\psi}\right) \right\|^{2}_{\widehat{\mathbf{R}}^{-1}} , \tag{16b}$$

with its gradient

$$\nabla \widehat{\mathcal{G}}(\boldsymbol{\psi}) = \widetilde{\mathbf{B}}^{-1} \cdot \left[\boldsymbol{\psi} - \overline{\boldsymbol{\psi}}^{b}\right] - \mathbf{W}^{T}_{\boldsymbol{\psi}^{(u)}} \cdot \widehat{\mathbf{R}}^{-1} \cdot \left[\mathbf{d} - \mathbf{W}_{\boldsymbol{\psi}^{(u)}} \cdot \boldsymbol{\psi}\right] \in \mathbb{R}^{n \times 1} , \tag{16c}$$

where $\mathbf{d} = \mathbf{o} - \mathcal{W}\left(\overline{\boldsymbol{\psi}}^{b}\right) \in \mathbb{R}^{m \times 1}$ is the innovation state on the observation $\mathbf{o}$. From (16c), the transition function

$$\mathbf{z}^{(u)} = \mathcal{K}\left(\boldsymbol{\psi}^{(u)}, \nabla \widehat{\mathcal{G}}\left(\boldsymbol{\psi}^{(u)}\right), \alpha\right) , \text{ with } \alpha \sim \mathcal{U}[0,\, 1] , \tag{17a}$$

is thus defined over samples, along the steepest descent direction of (16b):

$$\mathcal{K}\left(\boldsymbol{\psi}^{(u)}, \nabla \widehat{\mathcal{G}}\left(\boldsymbol{\psi}^{(u)}\right), \alpha\right) = \boldsymbol{\psi}^{(u)} - \alpha \cdot \nabla \widehat{\mathcal{G}}\left(\boldsymbol{\psi}^{(u)}\right) . \tag{17b}$$

Here, the uniform distribution $\mathcal{U}[0,\, 1]$ is drawn on $[0,\, 1]$. Hence, the acceptance/rejection rule, which is similar to that of the Tabu Search method, can be realized by

$$\boldsymbol{\psi}^{(u+1)} = \begin{cases} \boldsymbol{\psi}^{(u)}, & \text{for } \mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right) < \mathcal{G}\left(\mathbf{z}^{(u)}\right) , \\ \mathbf{z}^{(u)}, & \text{otherwise.} \end{cases} \tag{18}$$

The overall iterative process is then repeated for a fixed number of iterations, or until some predefined stopping criterion is met. Finally, the detailed TS-SGA can be summarized as the Algorithm 3.

Algorithm 3 The Tabu Search Single Gradient Approximation (TS-SGA) for Non-Gaussian Data Assimilation.

Require: Initial solution $\boldsymbol{\psi}^{(0)}$, typically $\boldsymbol{\psi}^{(0)} \leftarrow \overline{\boldsymbol{\psi}}^{b}$, and the maximum number of iterations U.

Ensure: A posterior mode approximation $\overline{\boldsymbol{\psi}}^{a}$ of Eq. (4).

for $u = 0 \rightarrow U$ **do**
  Linearize $\mathcal{W}(\boldsymbol{\psi})$ about $\boldsymbol{\psi}^{(u)}$ according to Eq. (16a).
  Compute the gradient $\widehat{\mathcal{G}}\left(\boldsymbol{\psi}^{(u)}\right)$ via Eq. (16c).
  Set $\alpha \sim \mathcal{U}[0,\, 1]$.
  Propose the state $\mathbf{z}^{(u)}$ by means of Eq. (17a).
  Set $\boldsymbol{\psi}^{(u+1)}$ as stated in Eq. (18).
  if stopping criterion is satisfied **then**
    break
Set $\overline{\boldsymbol{\psi}}^{a} \leftarrow \boldsymbol{\psi}^{(u)}$.

3.2. Tabu Search Multiple Gradient Approximation

With the gradient approximation (16c), we can generate a set of K random positive definite matrices:

$$\{\boldsymbol{\Pi}_1,\, \boldsymbol{\Pi}_2,\, \ldots,\, \boldsymbol{\Pi}_K\}\,, \tag{19}$$

where $\boldsymbol{\Pi}_k \in \mathbb{R}^{n\times n}$ (for $1 \leq k \leq K$) are used for generating a set of random directions:

$$\boldsymbol{\phi}^{(u,k)} = -\boldsymbol{\Pi}_k \cdot \nabla\widehat{\mathcal{G}}\left(\boldsymbol{\psi}^{(u)}\right) \in \mathbb{R}^{n\times 1}\,. \tag{20}$$

We can restrict the optimization problem (4) to the space spanned by such vectors:

$$\boldsymbol{\psi} = \boldsymbol{\psi}^{(u)} + \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu} \tag{21}$$

where $\boldsymbol{\mu} \in \mathbb{R}^{K\times 1}$ is a vector in redundant coordinates to be computed later. In addition, $\boldsymbol{\Phi}^{(u)}$ is given by

$$\boldsymbol{\Phi}^{(u)} = \left[\boldsymbol{\phi}^{(u,1)},\, \boldsymbol{\phi}^{(u,2)},\, \ldots,\, \boldsymbol{\phi}^{(u,K)}\right] \in \mathbb{R}^{n\times K}\,. \tag{22}$$

Substituting (21) into (16b), we have

$$\begin{aligned}\widehat{\mathcal{G}}\Big[\boldsymbol{\psi}^{(u)}+\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}\Big] &= \mathcal{Q}\left(\boldsymbol{\mu}\right)=\frac{1}{2}\cdot\left\|\boldsymbol{\delta\psi}^{(u)}+\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}\right\|^{2}_{\widetilde{\mathbf{B}}^{-1}}\\ &+ \frac{1}{2}\cdot\left\|\boldsymbol{\delta}\mathbf{y}^{(u)}-\mathbf{W}_{\boldsymbol{\psi}^{(u)}}\cdot\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}\right\|^{2}_{\widehat{\mathbf{R}}^{-1}}, \end{aligned} \tag{23}$$

where $\boldsymbol{\delta\psi}^{(u)}=\boldsymbol{\psi}^{(u)}-\overline{\boldsymbol{\psi}}^{b}\in\mathbb{R}^{n\times 1}$, and $\boldsymbol{\delta}\mathbf{y}^{(u)}=\mathbf{o}-\mathcal{W}\left(\boldsymbol{\psi}^{(u)}\right)\in\mathbb{R}^{m\times 1}$. Now the gradient of (23) becomes

$$\begin{aligned}\nabla\mathcal{Q}(\boldsymbol{\mu}) &= \left[\boldsymbol{\Phi}^{(u)}\right]^{T}\cdot\widetilde{\mathbf{B}}^{-1}\cdot\left[\boldsymbol{\delta\psi}^{(u)}+\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}\right]\\ &- \left[\mathbf{W}^{(u)}\right]^{T}\cdot\widehat{\mathbf{R}}^{-1}\left[\boldsymbol{\delta}\mathbf{y}^{(u)}-\mathbf{W}^{(u)}\cdot\boldsymbol{\mu}\right]\in\mathbb{R}^{K\times 1},\end{aligned}$$

where we have used $\mathbf{W}^{(u)}=\mathbf{W}_{\boldsymbol{\psi}^{(u)}}\cdot\boldsymbol{\Phi}^{(u)}\in\mathbb{R}^{m\times K}$ by setting this gradient to zero. As a result, the optimal weights can be computed by

$$\begin{aligned}\boldsymbol{\mu}^{*} &= -\left[\left[\boldsymbol{\Phi}^{(u)}\right]^{T}\cdot\widetilde{\mathbf{B}}^{-1}\cdot\boldsymbol{\Phi}^{(u)}+\left[\mathbf{W}^{(u)}\right]^{T}\cdot\widehat{\mathbf{R}}^{-1}\cdot\mathbf{W}^{(u)}\right]^{-1}\\ &\cdot\left[\left[\boldsymbol{\Phi}^{(u)}\right]^{T}\cdot\widetilde{\mathbf{B}}^{-1}\cdot\boldsymbol{\delta\psi}^{(u)}-\left[\mathbf{W}^{(u)}\right]^{T}\cdot\widehat{\mathbf{R}}^{-1}\cdot\boldsymbol{\delta}\mathbf{y}^{(u)}\right],\end{aligned} \tag{24}$$

over which our transition function

$$\mathbf{z}^{(u)}=\widehat{\mathcal{K}}\left(\boldsymbol{\psi}^{(u)},\,\boldsymbol{\Phi}^{(u)},\,\boldsymbol{\mu}^{*},\,\alpha\right),\ \text{with}\ \alpha\in\mathcal{U}[0,\,1], \tag{25}$$

can be expressed as

$$\widehat{\mathcal{K}}\left(\boldsymbol{\psi}^{(u)},\,\boldsymbol{\Phi}^{(u)},\,\boldsymbol{\mu}^{*},\,\alpha\right)=\boldsymbol{\psi}^{(u)}+\alpha\cdot\left[\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}^{*}\right]. \tag{26}$$

Here, the acceptance/rejection criteria is the same as (18). Again, the overall process is repeated until some predefined stopping criterion is met, often when a fixed number of maximum iterations is exceeded. In summary, the Tabu Search Multiple Gradient Approximations (TS-MGA) is detailed in the Algorithm 4.

3.3. Simulated Annealing Single Gradient Approximation and Simulated Annealing Multiple Gradient Approximations

The strict condition (18) can be difficult to satisfy, but it can be relaxed by using the well-known Metropolis Hastings criterion. Similar to TS-SGA, we can formulate a method such that, once a new state $\mathbf{z}^{(u)}$ is proposed, the acceptance/rejection criterion relies on the Boltzmann probability distribution:

$$\delta\left(\mathbf{z}^{(u)},\,\boldsymbol{\psi}^{(u)}\right)=\min\left\{1,\,\exp\left(-\left[\frac{\mathcal{G}\left(\mathbf{z}^{(u)}\right)-\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)}{T^{(u)}}\right]\right)\right\}, \tag{27}$$

Algorithm 4 The Tabu Search Multiple Gradient Approximations (TS-MGA) for Non-Gaussian Data Assimilation.

Require: Initial solution $\boldsymbol{\psi}^{(0)}$, typically $\boldsymbol{\psi}^{(0)} \leftarrow \overline{\boldsymbol{\psi}}^{b}$, the maximum number of iterations U.

Ensure: A posterior mode approximation $\overline{\boldsymbol{\psi}}^{a}$ of Eq. (4).

for $u = 0 \rightarrow U$ **do**
  Linearize $\mathcal{W}(\boldsymbol{\psi})$ about $\boldsymbol{\psi}^{(u)}$ according to (16a).
  Compute the gradient $\widehat{\mathcal{G}}\left(\boldsymbol{\psi}^{(u)}\right)$ via (16c).
  Compute the set of random matrices Eq. (19).
  Set $\boldsymbol{\Phi}^{(u)}$ as stated in (22).
  Calculate the optimal weights $\boldsymbol{\mu}^{*}$ via Eq. (24).
  Set $\alpha \sim \mathcal{U}[0,\, 1]$.
  Propose the state $\mathbf{z}^{(u)}$ by means of Eq. (25).
  Set $\boldsymbol{\psi}^{(u+1)}$ as stated in Eq. (18).
  if stopping criterion is satisfied **then**
    break
Set $\overline{\boldsymbol{\psi}}^{a} \leftarrow \boldsymbol{\psi}^{(u)}$.

where $T^{(u)} \in \mathbb{R}$ is the temperature at iteration u. Consequently, the current solution is updated as follows:

$$\boldsymbol{\psi}^{(u+1)} = \begin{cases} \mathbf{z}^{(u)}, & \text{with probability } \delta\left(\mathbf{z}^{(u)},\, \boldsymbol{\psi}^{(u)})\right), \\ \boldsymbol{\psi}^{(u)}, & \text{with probability } 1 - \delta\left(\mathbf{z}^{(u)},\, \boldsymbol{\psi}^{(u)})\right). \end{cases} \tag{28}$$

That is to say, the solutions with high-cost function values can be more likely to be accepted as long as $T^{(u)}$ is sufficiently large. For low temperature values, the acceptance/rejection criterion behaves similarly to that of TS based methods.

During iterations, the temperature $T^{(u)}$ is decreased by a so-called *cooling factor* ρ via a cooling schedule:

$$T^{(u)} = \rho \cdot T^{(u-1)}\,,$$

where ρ is typically in the range of [0.8, 0.95]. The Algorithm 5 details the *Simulated Annealing Single Gradient Approximations* (SA-SGA) steps. Obviously, in this context, a reduced-space approximation is also possible by constructing a set of surrogate basis vectors (22). The main idea is that the solution can be constrained to such sub-spaces whose dimensions can be much less than those of actual search spaces; once a solution is found, it is projected back onto the actual space of feasible solutions. This strategy can be employed so as to reduce the computational complexity of the SA-SGA formulation during iterations. We can now call this initiative the *Simulated Annealing Multiple Gradient Approximations* (SA-MGA), and its steps are summarized in the Algorithm 6.

Algorithm 5 The Simulated Annealing Single Gradient Approximation (SA-SGA) for Non-Gaussian Data Assimilation.

Require: Initial solution $\boldsymbol{\psi}^{(0)}$, typically $\boldsymbol{\psi}^{(0)} \leftarrow \overline{\boldsymbol{\psi}}^{b}$, initial temperature T_{ini}, cooler factor ρ, and the lowest temperature $T_{\min}$.
Ensure: A posterior mode approximation $\overline{\boldsymbol{\psi}}^{a}$ of Eq. (4).

$T^{(0)} \leftarrow T_{\text{ini}}$
$u \leftarrow 0$
while $T^{(u)} > T_{\min}$ **do**
    Linearize $\mathcal{W}(\boldsymbol{\psi})$ about $\boldsymbol{\psi}^{(u)}$ according to Eq. (16a).
    Compute the gradient $\widehat{\mathcal{G}}\left(\boldsymbol{\psi}^{(u)}\right)$ via Eq. (16c).
    Set $\alpha \sim \mathcal{U}[0, 1]$.
    Propose the state $\mathbf{z}^{(u)}$ by means of Eq. (17a).
    Set $\boldsymbol{\psi}^{(u+1)}$ as stated in Eq. (28).
    if stopping criterion is satisfied **then**
        break
    $u \leftarrow u + 1$
    $T^{(u)} \leftarrow \rho \cdot T^{(u-1)}$
Set $\overline{\boldsymbol{\psi}}^{a} \leftarrow \boldsymbol{\psi}^{(u)}$.

Algorithm 6 The Simulated Annealing Multiple Gradient Approximations (SA-MGA) for Non-Gaussian Data Assimilation.

Require: Initial solution $\boldsymbol{\psi}^{(0)}$, typically $\boldsymbol{\psi}^{(0)} \leftarrow \overline{\boldsymbol{\psi}}^{b}$, initial temperature T_{ini}, cooler factor ρ, and the lowest temperature $T_{\min}$.
Ensure: A posterior mode approximation $\overline{\boldsymbol{\psi}}^{a}$ of Eq. (4).

$T^{(0)} \leftarrow T_{\text{ini}}$
$u \leftarrow 0$
while $T^{(u)} > T_{\min}$ **do**
    Linearize $\mathcal{W}(\boldsymbol{\psi})$ about $\boldsymbol{\psi}^{(u)}$ according to Eq. (16a).
    Compute the gradient $\widehat{\mathcal{G}}\left(\boldsymbol{\psi}^{(u)}\right)$ via Eq. (16c).
    Compute the set of random matrices Eq. (19).
    Set $\boldsymbol{\Phi}^{(u)}$ as stated in Eq. (22).
    Calculate the optimal weights $\boldsymbol{\mu}^{*}$ via Eq. (24).
    Set $\alpha \sim \mathcal{U}[0, 1]$.
    Propose the state $\mathbf{z}^{(u)}$ by means of Eq. (25).
    Set $\boldsymbol{\psi}^{(u+1)}$ as stated in Eq. (28).
    if stopping criterion is satisfied **then**
        break
    $u \leftarrow u + 1$
    $T^{(u)} \leftarrow \rho \cdot T^{(u-1)}$
Set $\overline{\boldsymbol{\psi}}^{a} \leftarrow \boldsymbol{\psi}^{(u)}$.

3.4. Building the Posterior Ensemble

Once the optimization process is completed, the obtained solution $\boldsymbol{\psi}^{(u)}$ serves as the analysis mean about which the posterior members are built by means of the Posterior Ensemble Kalman Filter (P-EnKF) equations [20]. Now the e-th posterior member is estimated as follows:

$$\boldsymbol{\psi}^{a(e)} = \overline{\boldsymbol{\psi}}^{a} + \boldsymbol{\delta\psi}^{a(e)}\,,\ \text{for } 1 \leq e \leq N\,,$$

where $\boldsymbol{\delta\psi}^{a(e)} \in \mathbb{R}^{n\times 1}$ follows the distribution

$$\boldsymbol{\delta\psi}^{a(e)} \sim \mathcal{N}\left(\mathbf{0},\ \left[\widehat{\mathbf{L}}^{T}\cdot\widehat{\mathbf{D}}^{-1}\cdot\widehat{\mathbf{L}}\right]^{-1}\right)\,. \tag{29}$$

The estimate of the posterior precision covariance matrix can be done via a modified Cholesky decomposition

$$\widehat{\mathbf{L}}^{T}\cdot\widehat{\mathbf{D}}^{-1}\cdot\widehat{\mathbf{L}} = \widetilde{\mathbf{B}}^{-1} + \mathbf{W}^{T}_{\overline{\boldsymbol{\psi}}^{a}}\cdot\widehat{\mathbf{R}}^{-1}\cdot\mathbf{W}_{\overline{\boldsymbol{\psi}}^{a}} \in \mathbb{R}^{n\times n}\,.$$

By using the formulation [19], the matrix inversion in (29) is not actually needed. Once all prior members are updated, the analysis ensemble is propagated in time until new observations become available:

$$\boldsymbol{\psi}^{b[e]}_{\ell} = \mathcal{M}_{t_{\ell-1}\to t_{\ell}}\left(\boldsymbol{\psi}^{a[e]}_{\ell-1}\right)\,,\ \text{for } 1 \leq e \leq M\,,$$

for all $1 \leq \ell \leq M$ where M is the number of observations inside the current assimilation window.

3.5. Convergence Analysis of Proposed Methods

In order to prove the convergence of the TS-MGA, we now consider the assumptions (A), (B), and the condition

$$\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)^{T}\cdot\boldsymbol{\phi}^{(u,k)} < 0,\ \text{ for } 1 \leq k \leq K\,. \tag{30}$$

With the above assumptions, global convergence for the TS-MGA method can be ensured by the next theorem with the necessary conditions.

Theorem 1. *If* (A), (B), *and* (30) *hold, the TS-MGA with random line search generates an infinite sequence* $\left\{\boldsymbol{\psi}^{(u)}\right\}_{u=0}^{\infty}$, *then*

$$\lim_{u\to\infty}\left[\frac{-\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)^{T}\cdot\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}^{*}}{\left\|\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}^{*}\right\|}\right]^{2} = 0 \tag{31}$$

holds.

Proof. From Taylor series, the acceptance condition (9), and the Mean Value Theorem, we know

$$\begin{aligned}\mathcal{G}\left(\boldsymbol{\psi}^{(u)}+\alpha^*\cdot\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}^*\right) &= \mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)\\ &+ \alpha^*\cdot\int_0^1\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}+\alpha^*\cdot t\cdot\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}^*\right)^T\\ &\cdot \boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}^*\cdot dt\,,\end{aligned}$$

where α^* is given by (15). Then, we also have

$$\begin{aligned}\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)-\mathcal{G}\left(\boldsymbol{\psi}^{(u+1)}\right) &\geq -\alpha^*\cdot\int_0^1\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}+\alpha^*\cdot t\cdot\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}^*\right)^T\\ &\cdot \boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}^*\cdot dt\end{aligned}$$

for any $\boldsymbol{\psi}^{(u+1)}$ on the direction $\boldsymbol{\psi}^{(u)}+\alpha\cdot\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}^*$ (with $\rho\in[0,\,1]$). Thus, we get

$$\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)-\mathcal{G}\left(\boldsymbol{\psi}^{(u+1)}\right)\geq\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)-\mathcal{G}\left(\boldsymbol{\psi}^{(u)}+\alpha^*\cdot\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}^*\right),$$

so that

$$\begin{aligned}\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right) &- \mathcal{G}\left(\boldsymbol{\psi}^{(u+1)}\right)\geq-\alpha^*\cdot\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)^T\cdot\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}^*\\ &- \alpha^*\cdot\int_0^1\left[\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}+\alpha^*\cdot t\cdot\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}^*\right)-\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)\right]^T\\ &\cdot \boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}^*\cdot dt\,.\end{aligned}$$

Using the Cauchy Schwarz inequality, we have

$$\begin{aligned}
\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right) &- \mathcal{G}\left(\boldsymbol{\psi}^{(u+1)}\right) \geq -\alpha^* \cdot \nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)^T \cdot \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^* \\
&- \alpha^* \cdot \int_0^1 \left\| \nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)} + \alpha^* \cdot t \cdot \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^*\right) - \nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right) \right\| \\
&\cdot \left\| \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^* \right\| \cdot dt \\
&\geq -\alpha^* \cdot \nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)^T \cdot \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^* \\
&- \alpha^* \cdot \int_0^1 L \cdot \left\| \alpha^* \cdot t \cdot \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^* \right\| \cdot \left\| \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^* \right\| \cdot dt \\
&= -\alpha^* \cdot \nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)^T \cdot \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^* \\
&- \alpha^* \cdot L \cdot \left\| \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^* \right\| \cdot \int_0^1 \left\| t \cdot \alpha^* \cdot \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^* \right\| \cdot dt \\
&= -\alpha^* \cdot \nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)^T \cdot \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^* - \frac{1}{2} \cdot \alpha^{*2} \cdot L \cdot \left\| \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^* \right\|^2 ,
\end{aligned}$$

to ensure decrease of (3), we choose alpha as

$$\alpha^* = -\frac{\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)^T \cdot \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^*}{L \cdot \left\| \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^* \right\|^2} ,$$

leading to

$$\begin{aligned}
\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right) &- \mathcal{G}\left(\boldsymbol{\psi}^{(u+1)}\right) \geq \frac{\left[\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)^T \cdot \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^*\right]^2}{L \cdot \left\| \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^* \right\|^2} \\
&- \frac{1}{2} \cdot \frac{\left[-\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)^T \cdot \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^*\right]^2}{L \cdot \left\| \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^* \right\|^2} \\
&= \frac{1}{2 \cdot L} \cdot \left[-\frac{\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)^T \cdot \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^*}{\left\| \boldsymbol{\Phi}^{(u)} \cdot \boldsymbol{\mu}^* \right\|} \right]^2 .
\end{aligned}$$

By (A) and (30), it is straightforward to show that $\left\{\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)\right\}_{u=0}^{\infty}$ is a monotonically decreasing number sequence with a bound below. Therefore, $\left\{\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)\right\}_{u=0}^{\infty}$ has a limit, and consequently (31) holds. □

It is worth pointing out that the TS-SGA is a particular case of the TS-MGA when the search direction is $-\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)$. Thus, the descent condition

$$-\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)^{T}\cdot\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)=-\left\|\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)\right\|^{2}<0 \tag{32}$$

is always satisfied. Simiarly, the next Theorem states the necessary conditions for guaranteeing the convergence of the TS-SGA.

Theorem 2. *If* (A), (B), *and* (32) *hold, the TS-SGA with random line search generates an infinite sequence* $\left\{\boldsymbol{\psi}^{(u)}\right\}_{u=0}^{\infty}$, *then*

$$\lim_{u\to\infty}\left[\frac{-\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)^{T}\cdot\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)}{\left\|\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)\right\|}\right]^{2}=0 \tag{33}$$

holds.

Proof. The out of proving this Theorem can be done in a similar way to that of proving Theorem 1. It is enough to note that the search direction $\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}^{*}$ is now replaced by $-\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)$ and making use of (32). Then, the results in the theorem follow. □

Now we can state the the convergence of SA-SGA and SA-MGA in the following two corollaries 1 and 2, respectively. It should be noted that, as $T^{(u)}$ goes to 0, the acceptance/rejection rule of SA methods is similar to that of TS algorithms.

Corollary 1. *If* (A), (B), *and* (32) *are true, as* $T^{(u)}\to 0$, *the SA-SGA with random line search generates an infinite sequence* $\left\{\boldsymbol{\psi}^{(u)}\right\}_{u=0}^{\infty}$, *then*

$$\lim_{u\to\infty}\left[\frac{-\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)^{T}\cdot\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)}{\left\|\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)\right\|}\right]^{2}=0$$

holds.

Corollary 2. *If* (A), (B), *and* (30) *are true, as* $T^{(u)}\to 0$, *the TS-MGA with random line search generates an infinite sequence* $\left\{\boldsymbol{\psi}^{(u)}\right\}_{u=0}^{\infty}$, *then*

$$\lim_{u\to\infty}\left[\frac{-\nabla\mathcal{G}\left(\boldsymbol{\psi}^{(u)}\right)^{T}\cdot\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}^{*}}{\left\|\boldsymbol{\Phi}^{(u)}\cdot\boldsymbol{\mu}^{*}\right\|}\right]^{2}=0$$

holds.

We now are ready to perform some numerical experiments to show that the proposed approaches can indeed work well.

4. Numerical Experiments and Results

We now test and validate our proposed methods by using seven different statistical models. By using Bayes' rule, we know that the posterior error distribution reads:

$$\mathcal{P}\left(\boldsymbol{\psi}|\mathbf{o}\right) \propto \exp\left(-\mathcal{G}(\boldsymbol{\psi})\right) , \tag{34}$$

where $\mathcal{G}(\boldsymbol{\psi})$ is given in (3). We consider the nonlinear observation operator [65]:

$$\mathcal{W}\left(\boldsymbol{\psi}\right) \equiv \left\{\mathcal{W}\left(\boldsymbol{\psi}\right)\right\}_j = \frac{\left\{\boldsymbol{\psi}\right\}_j}{2} \cdot \left[1 + \left(\frac{|\left\{\boldsymbol{\psi}\right\}_j|}{2}\right)^{\gamma-1}\right] , \tag{35}$$

where j corresponds to the j-th observed component, for $1 \leq j \leq m$. The values of γ vary in $1 \leq \gamma \leq 7$ from which seven different statistical models in (34) are obtained, some of these models can be seen in figure (2). Thus, for each value of γ, a different optimization problem of the form:

$$\boldsymbol{\psi}^a = \arg\max_{\boldsymbol{\psi}} \mathcal{P}\left(\boldsymbol{\psi}|\mathbf{o}\right) ,$$

is derived. The experimental settings are as follows:

- We make use the Lorenz 96 model as our surrogate numerical model [66] from which samples from prior error distributions are obtained. This model is defined over a set of nonlinear ordinary differential equations:

$$\frac{dx_j}{dt} = \begin{cases} (x_2 - x_{n-1}) \cdot x_n - x_1 + F & \text{for } i = 1, \\ (x_{i+1} - x_{i-2}) \cdot x_{i-1} - x_i + F & \text{for } 2 \leq i \leq n-1, \\ (x_1 - x_{n-2}) \cdot x_{n-1} - x_n + F & \text{for } i = n, \end{cases} \tag{36}$$

  where x_i is the i-th model component (for $1 \leq i \leq n$). Each model component corresponds to a particle which fluctuates in the atmosphere and exhibits some properties such as advection and internal dissipation [67]. Besides, the Lorenz 96 model exhibits chaotic behavior when the external force F is set to eight, which makes the model attractive for testing emerging data assimilation schemes.

- No model errors are considered during the experiments.

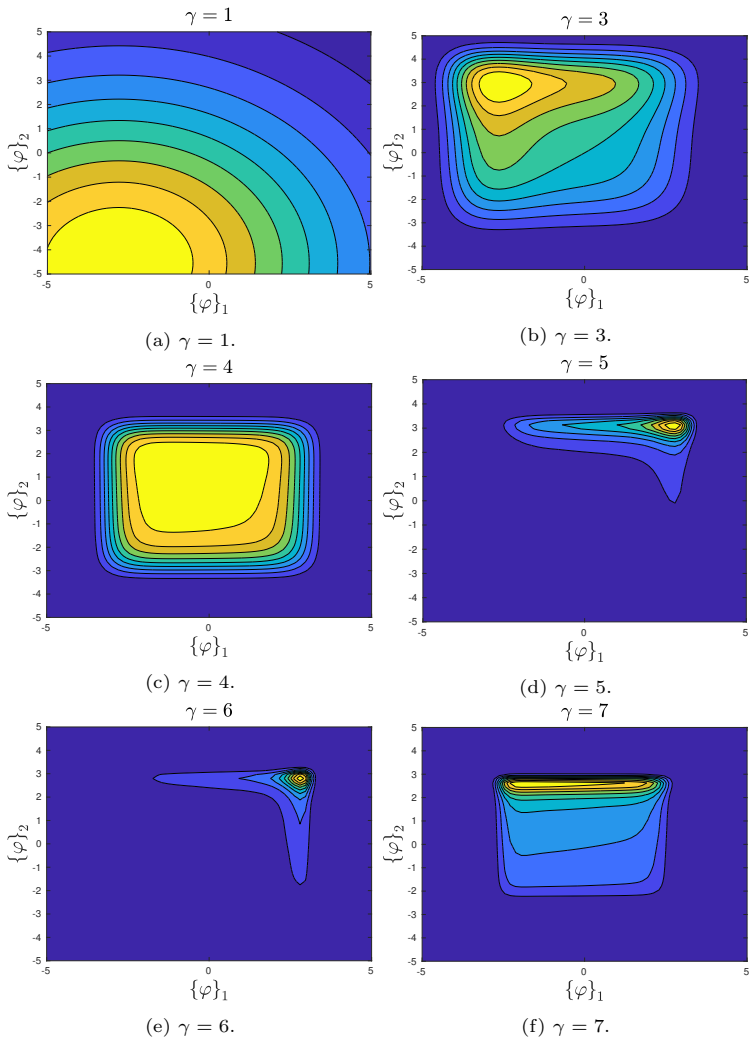
- The number of model components n is set as $n = 40$.

- The propagation of an initial perturbed state is carried out over a long time period so as to be consistent with the model (36). As a result, the actual initial solution $\boldsymbol{\psi}_0^*$ is obtained. Similar operations are applied for

355 building the initial background state as well as the initial background ensemble.

- For the background ensemble, we create a pool of 10^5 members in the experiments. Random members are sampled from such pool to obtain initial background ensembles.

360 - We consider two observational grids, in the first case, the number of observed components p is set to 70%, while in the last one is set to 90%. Note that, $m = p \cdot n$.



(a) $\gamma = 1$. (b) $\gamma = 3$. (c) $\gamma = 4$. (d) $\gamma = 5$. (e) $\gamma = 6$. (f) $\gamma = 7$.

Figure 2: Two dimensional projections of likelihood functions (data error distributions) for different values of γ. Seven different statistical models are tried during the experiments.

- The assimilation window consists of $M = 20$ observations that are evenly spaced. Observations are available every 17 hours. They are synthetically built by using the probability distribution:

$$\mathbf{o}_\ell \sim \mathcal{N}\left(\mathcal{W}\left(\boldsymbol{\psi}_\ell^*\right), \widehat{\mathbf{R}}\right), \text{ for } 1 \leq \ell \leq M, \tag{37}$$

where the covariance matrix $\widehat{\mathbf{R}}$ of the data-errors is diagonal with diagonal elements being $\sigma^2 = 0.01^2$. This essentially mimics the realistic behaviour of observations when collected via sensors.

Now the parameter settings are as follows:

- The ensemble size $N = 20$.
- For the Tabu Search (TS) based methods, the number of maximum iterations varies in $U \in \{100, 200, 300\}$.
- For the Simulated Annealing (SA) based methods, the cooling factor ρ is set to be in $\rho \in \{0.85, 0.90, 0.95\}$.
- The sub-space approximations of TS and SA use spaces of sizes $K \in \{10, 20, 30, 40\}$.
- We consider the L_2-norm of errors in order to estimate the actual error at the different assimilation steps ℓ, for $1 \leq \ell \leq M$,

$$\lambda_\ell = \left\| \boldsymbol{\psi}_\ell^* - \overline{\boldsymbol{\psi}}_\ell^a \right\|_2 = \sqrt{[\boldsymbol{\psi}_\ell^* - \boldsymbol{\psi}_\ell^a]^T \cdot [\boldsymbol{\psi}_\ell^* - \boldsymbol{\psi}_\ell^a]}, \tag{38}$$

where $\boldsymbol{\psi}_\ell^*$ and $\boldsymbol{\psi}_\ell^a$ are the reference solutions and the solutions from the analysis, respectively.

- On average, the errors over a given assimilation window are measured by using the Root-Mean-Square-Error (RMSE):

$$\lambda = \sqrt{\frac{1}{M} \cdot \sum_{\ell=1}^{M} \lambda_\ell^2}. \tag{39}$$

- For each parameter setting of γ and p, 10 independent runs are performed for each method so as to assess the averaged accuracy of the proposed methods by means of the metrics (38) and (39).

For a complete assimilation window, the averages of error norms are shown in Figs 3 and 4 for the TS-SGA and the SA-SGA formulations, respectively. For the TS-SGA, a different number of iterations U are attempted, while different cooling factors ρ are also used for the SA-SGA implementation. It can be seen clearly that, in both cases, the error norms decrease as the ensemble moves forward in the assimilation window as expected. As more data or information

is assimilated into the actual imperfect model, the uncertainties will be largely reduced during assimilation steps.

On the other hand, for a large number of observed components, highly nonlinear observation operators can be less sensitive to overfitting during assimilation stages. In all cases, the behavior of both methods is similar, regardless of their parameter configurations. This may be mainly attributed to the fact that the sampling procedure is performed along a gradient approximation of the 3D-Var cost function. Therefore, high-quality states can be obtained from such set of directions. In addition, the parameter configurations (U or ρ where appropriate) do not influence much on the quality of solutions; this feature is attractive since those parameters can be hard to tune in practice. For example, the cooling factor ρ in SA based methods and the number of iterations U in TS inspired formulations can be considered as hyper-parameters, thus any insensitivity to such parameters can be desirable.

As we briefly mentioned before, the computational efforts of SGA formulations can be decreased by using sub-space approximations during optimization steps. The results for the TS-MGA and the SA-MGA, respectively, can be seen in Figs 4 and 6 where $p = 70\%$ of model components are observed from the model state. Again, for all configurations and parameter settings, the proposed methods can reduce initial background errors as observations are gradually used and assimilated. Furthermore, reduced-space approximations in some cases can provide results similar to those of full-space approximations. For $K = 10$, it can be seen clearly that the performance of MGA based methods can degrade for highly nonlinear observation operators (i.e., $\gamma = 7$), though this is a reasonable accuracy considering the trade-off between the computational effort of computing steps in such sub-spaces. However, in terms of RMSE values, for different values of γ, all methods behave similarly as can be seen in the Tables 1 and 2 for $p = 70\%$, and in the Tables 3 and 4 for $p = 90\%$. The results are reported after removing the spin-off period (the first 6 assimilation steps) to better understand the behavior of filters once observations have been injected into the numerical model. Note that, such assimilation steps can be performed within a reasonable computational time, for instance, posterior states computations are bounded by seconds as can be seen in the Tables 5 and 6 for different values of γ.

In figures 7 and 8, we report some results of a single assimilation step for the TS-MGA and the SA-MGA, respectively. We consider the initial assimilation step since no information from the actual system dynamics (36) has been injected into the numerical forecast. The results are shown in the logarithm scale for the cost function values and the optimization step. As can be seen, in both cases, as the sub-spaces dimensions are increased, the methods can converge faster to posterior modes of the error distribution. This is more evident for TS based methods, for SA inspired algorithms equivalent results can be obtained in a similar number of iterations but, it is evident that the more degrees of freedom (sub-spaces dimensions) the faster their convergence. Note that, some fluctuations in cost function values among iterations can be observed for the SA-MGA, this can be possible owing to the acceptance/rejection rule of such method wherein solutions with large cost function values can be considered over

short time periods to avoid getting trap in local minimizers. Besides, the acceptance/rejection rule of SA methods can exploit sub-spaces dimensions by providing more accurate results as those are increased.

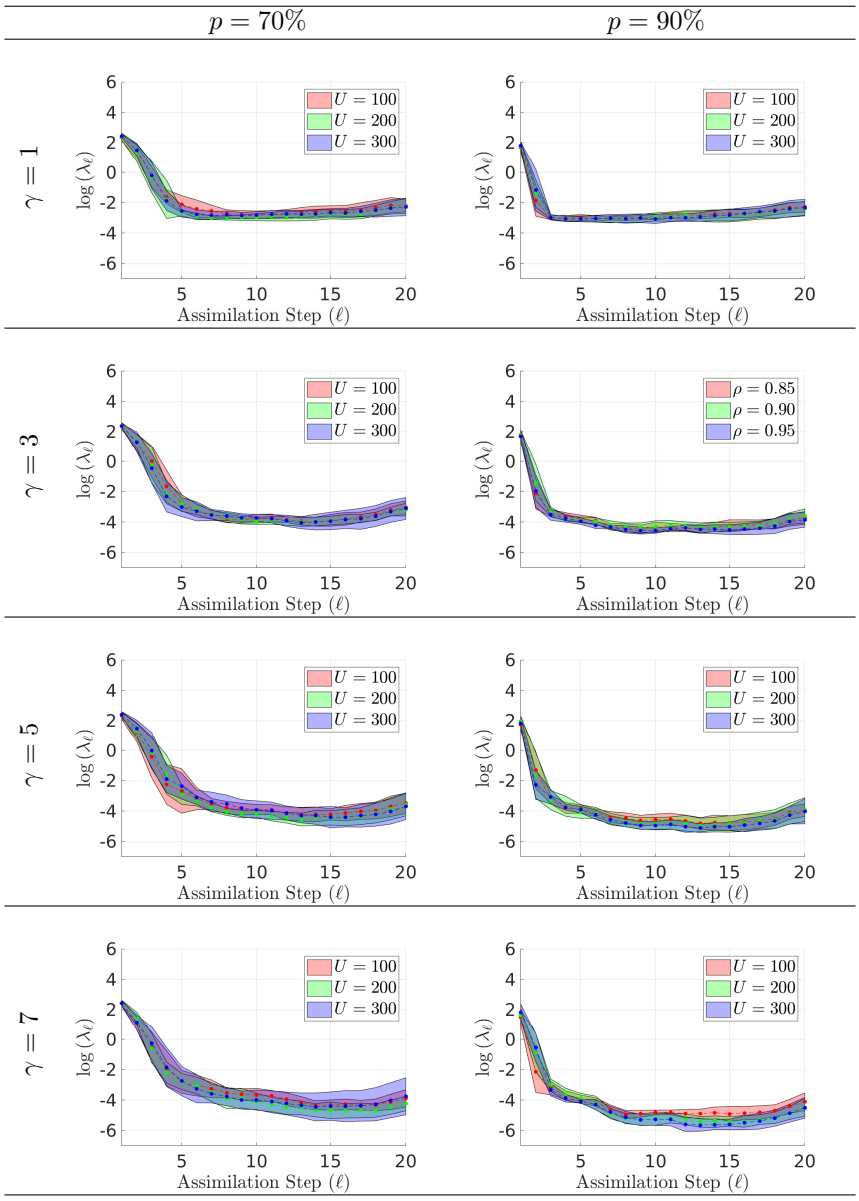


Figure 3: Averages (dashed lines) and standard deviations (shaded regions) of error norms for the TS-SGA implementation, different values of parameters U, and $p = 70\%$ of components observed from the model state.

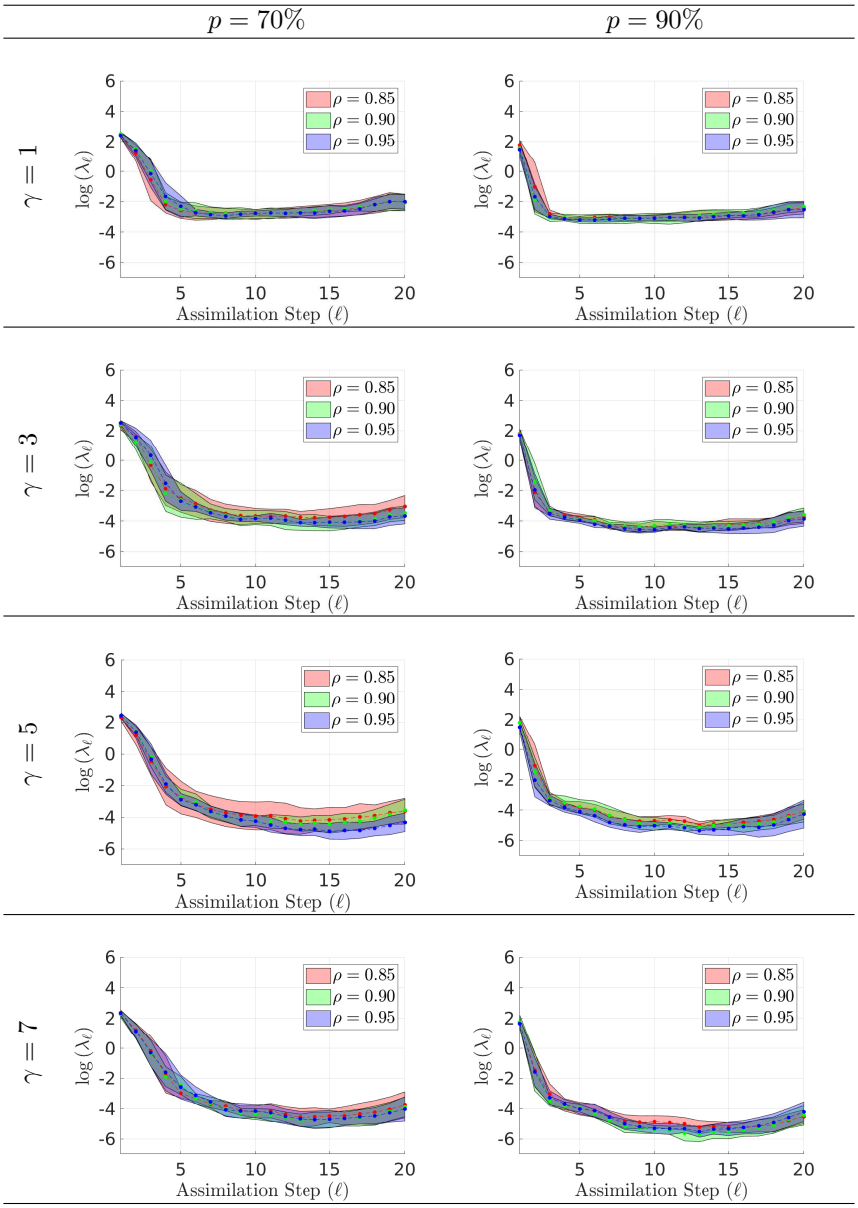


Figure 4: Averages (dashed lines) and standard deviations (shaded regions) of error norms for the SA-SGA implementation, different values of parameters ρ, and $p = 70\%$ of components observed from the model state.

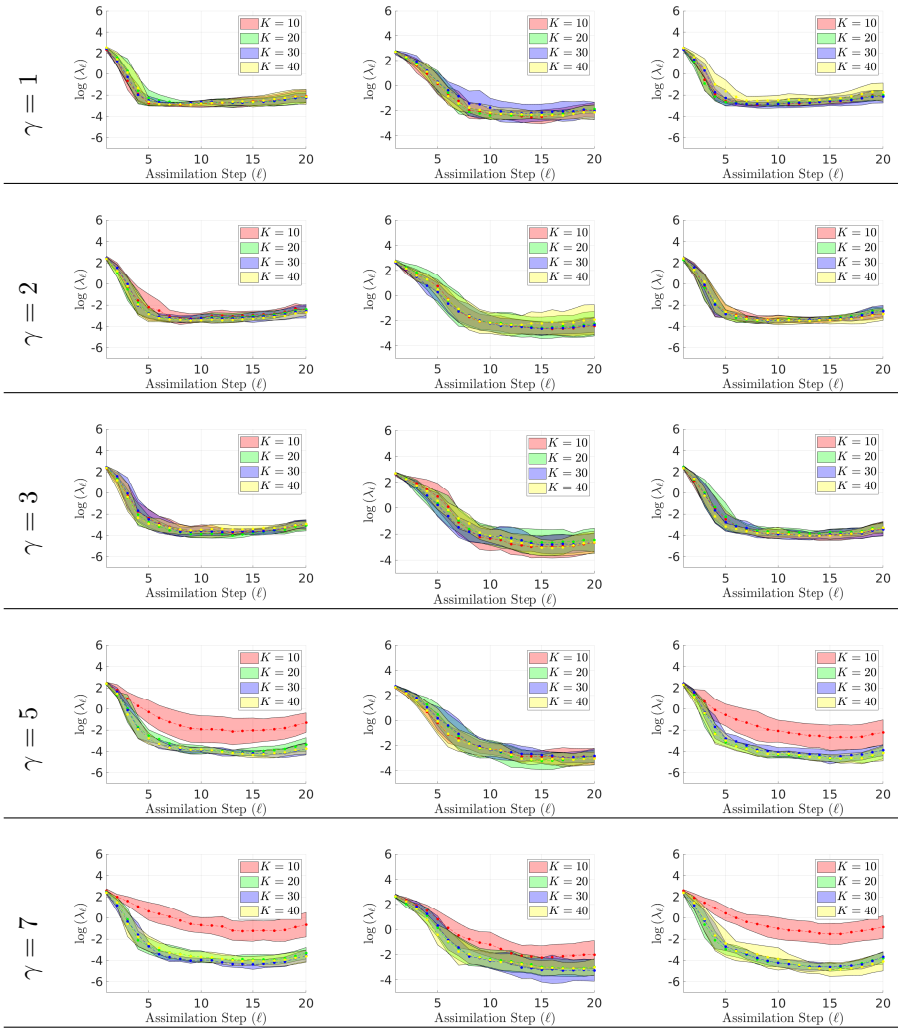


Figure 5: Averages and standard deviations of error norms for the SA-MGA implementation, different values of parameter U, different sub-space sizes K, and $p = 70\%$ of components observed from the model state.

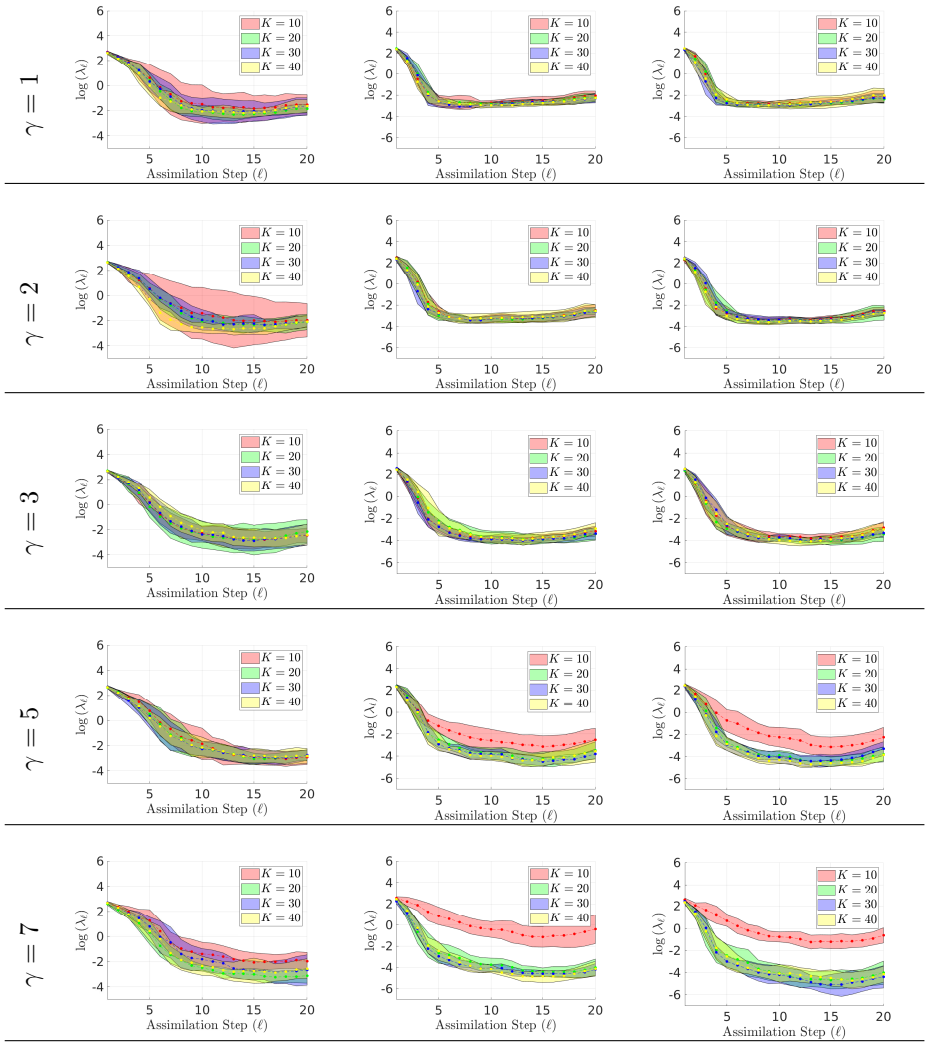


Figure 6: Averages and standard deviations of error norms for the SA-MGA implementation, different values of parameter ρ, different sub-space sizes K, and $p = 70\%$ of components observed from the model state.

	Tabu Search Based Methods				Simulated Annealing Based Methods			
γ	U	K	TS-MGA	TS-SGA	ρ	K	SA-MGA	SA-SGA
1	100	10	0.088	0.092	0.85	10	0.089	0.095
		20	0.087			20	0.077	
		30	0.080			30	0.070	
		40	0.090			40	0.091	
	200	10	0.097		0.90	10	0.099	
		20	0.078			20	0.081	
		30	0.065			30	0.073	
		40	0.068			40	0.072	
	300	10	0.085		0.95	10	0.087	
		20	0.088			20	0.071	
		30	0.090			30	0.074	
		40	0.152			40	0.097	
2	100	10	0.080	0.048	0.85	10	0.046	0.062
		20	0.056			20	0.046	
		30	0.055			30	0.042	
		40	0.044			40	0.049	
	200	10	0.043		0.90	10	0.055	
		20	0.061			20	0.049	
		30	0.041			30	0.050	
		40	0.065			40	0.061	
	300	10	0.047		0.95	10	0.050	
		20	0.049			20	0.050	
		30	0.049			30	0.045	
		40	0.046			40	0.040	
3	100	10	0.041	0.031	0.85	10	0.030	0.045
		20	0.032			20	0.034	
		30	0.042			30	0.040	
		40	0.041			40	0.035	
	200	10	0.026		0.90	10	0.031	
		20	0.035			20	0.039	
		30	0.026			30	0.028	
		40	0.028			40	0.043	
	300	10	0.032		0.95	10	0.039	
		20	0.034			20	0.030	
		30	0.031			30	0.029	
		40	0.028			40	0.036	

Table 1: Averages of Root-Mean-Square-Errors (RMSE) across an assimilation window with 20 observations for 10 repetitions. The non-linear term γ ranges in $\gamma \in \{1, 2, 3\}$, likewise $p = 70\%$.

	Tabu Search Based Methods				Simulated Annealing Based Methods			
γ	U	K	TS-MGA	TS-SGA	ρ	K	SA-MGA	SA-SGA
5	100	10	0.028	0.023	0.85	10	0.023	0.031
		20	0.022			20	0.020	
		30	0.024			30	0.016	
		40	0.024			40	0.024	
	200	10	0.029		0.90	10	0.019	
		20	0.015			20	0.021	
		30	0.020			30	0.015	
		40	0.020			40	0.019	
	300	10	0.017		0.95	10	0.015	
		20	0.019			20	0.012	
		30	0.018			30	0.026	
		40	0.020			40	0.017	
6	100	10	0.027	0.020	0.85	10	0.027	0.021
		20	0.028			20	0.019	
		30	0.023			30	0.016	
		40	0.021			40	0.021	
	200	10	0.022		0.90	10	0.027	
		20	0.015			20	0.018	
		30	0.018			30	0.018	
		40	0.017			40	0.020	
	300	10	0.019		0.95	10	0.016	
		20	0.018			20	0.017	
		30	0.015			30	0.021	
		40	0.012			40	0.022	
7	100	10	0.063	0.020	0.85	10	0.060	0.019
		20	0.020			20	0.013	
		30	0.024			30	0.017	
		40	0.023			40	0.017	
	200	10	0.078		0.90	10	0.027	
		20	0.016			20	0.015	
		30	0.016			30	0.014	
		40	0.018			40	0.016	
	300	10	0.026		0.95	10	0.016	
		20	0.015			20	0.014	
		30	0.012			30	0.016	
		40	0.014			40	0.016	

Table 2: Averages of Root-Mean-Square-Errors (RMSE) across an assimilation window with 20 observations for 10 repetitions. The non-linear term γ ranges in $\gamma \in \{5, 6, 7\}$, likewise $p = 70\%$.

	Tabu Search Based Methods				Simulated Annealing Based Methods			
γ	U	K	TS-MGA	TS-SGA	ρ	K	SA-MGA	SA-SGA
1	100	10	0.106	0.076	0.85	10	0.086	0.064
		20	0.099			20	0.084	
		30	0.083			30	0.090	
		40	0.081			40	0.078	
	200	10	0.084		0.90	10	0.090	
		20	0.088			20	0.070	
		30	0.093			30	0.103	
		40	0.099			40	0.073	
	300	10	0.089		0.95	10	0.074	
		20	0.065			20	0.070	
		30	0.079			30	0.074	
		40	0.083			40	0.079	
2	100	10	0.052	0.053	0.85	10	0.041	0.043
		20	0.060			20	0.042	
		30	0.043			30	0.039	
		40	0.039			40	0.039	
	200	10	0.048		0.90	10	0.043	
		20	0.037			20	0.039	
		30	0.051			30	0.037	
		40	0.039			40	0.034	
	300	10	0.037		0.95	10	0.035	
		20	0.041			20	0.038	
		30	0.035			30	0.042	
		40	0.051			40	0.036	
3	100	10	0.039	0.027	0.85	10	0.020	0.017
		20	0.022			20	0.021	
		30	0.028			30	0.026	
		40	0.027			40	0.024	
	200	10	0.026		0.90	10	0.018	
		20	0.019			20	0.018	
		30	0.018			30	0.021	
		40	0.020			40	0.019	
	300	10	0.019		0.95	10	0.021	
		20	0.020			20	0.015	
		30	0.017			30	0.014	
		40	0.020			40	0.018	

Table 3: Averages of Root-Mean-Square-Errors (RMSE) across an assimilation window with 20 observations for 10 repetitions. The non-linear term γ ranges in $\gamma \in \{1, 2, 3\}$, likewise $p = 90\%$.

	Tabu Search Based Methods				Simulated Annealing Based Methods			
γ	U	K	TS-MGA	TS-SGA	ρ	K	SA-MGA	SA-SGA
5	100	10	0.219	0.013	0.85	10	0.162	0.011
		20	0.034			20	0.014	
		30	0.016			30	0.014	
		40	0.014			40	0.013	
	200	10	0.139		0.90	10	0.266	
		20	0.018			20	0.021	
		30	0.009			30	0.007	
		40	0.010			40	0.011	
	300	10	0.166		0.95	10	0.247	
		20	0.011			20	0.010	
		30	0.011			30	0.008	
		40	0.011			40	0.010	
6	100	10	0.370	0.013	0.85	10	0.382	0.010
		20	0.047			20	0.046	
		30	0.013			30	0.011	
		40	0.014			40	0.009	
	200	10	0.349		0.90	10	0.323	
		20	0.031			20	0.019	
		30	0.011			30	0.010	
		40	0.010			40	0.008	
	300	10	0.583		0.95	10	0.270	
		20	0.022			20	0.033	
		30	0.012			30	0.007	
		40	0.010			40	0.009	
7	100	10	0.639	0.011	0.85	10	0.477	0.008
		20	0.167			20	0.167	
		30	0.014			30	0.012	
		40	0.012			40	0.008	
	200	10	0.942		0.90	10	0.375	
		20	0.135			20	0.056	
		30	0.008			30	0.009	
		40	0.011			40	0.006	
	300	10	0.495		0.95	10	0.413	
		20	0.070			20	0.153	
		30	0.011			30	0.007	
		40	0.009			40	0.009	

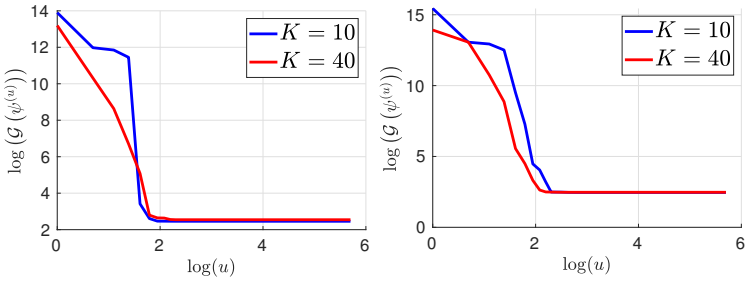
Table 4: Averages of Root-Mean-Square-Errors (RMSE) across an assimilation window with 20 observations for 10 repetitions. The non-linear term γ ranges in $\gamma \in \{5, 6, 7\}$, likewise $p = 90\%$.

	Tabu Search Based Methods				Simulated Annealing Based Methods			
γ	U	K	TS-MGA	TS-SGA	ρ	K	SA-MGA	SA-SGA
1	100	10	0.112	0.092	0.85	10	0.081	0.103
		20	0.091			20	0.099	
		30	0.082			30	0.110	
		40	0.091			40	0.112	
	200	10	0.084		0.90	10	0.066	
		20	0.101			20	0.082	
		30	0.101			30	0.087	
		40	0.095			40	0.086	
	300	10	0.076		0.95	10	0.086	
		20	0.076			20	0.086	
		30	0.091			30	0.086	
		40	0.095			40	0.092	
2	100	10	0.053	0.047	0.85	10	0.049	0.064
		20	0.063			20	0.046	
		30	0.052			30	0.046	
		40	0.044			40	0.051	
	200	10	0.047		0.90	10	0.057	
		20	0.041			20	0.045	
		30	0.044			30	0.055	
		40	0.038			40	0.065	
	300	10	0.047		0.95	10	0.050	
		20	0.055			20	0.045	
		30	0.045			30	0.039	
		40	0.053			40	0.047	
3	100	10	0.032	0.030	0.85	10	0.036	0.036
		20	0.031			20	0.028	
		30	0.027			30	0.034	
		40	0.034			40	0.032	
	200	10	0.032		0.90	10	0.034	
		20	0.026			20	0.022	
		30	0.023			30	0.027	
		40	0.023			40	0.025	
	300	10	0.027		0.95	10	0.039	
		20	0.025			20	0.029	
		30	0.032			30	0.024	
		40	0.025			40	0.032	

Table 5: Average of elapsed times, in seconds, for the compared methods in a single assimilation step, the number of repetition reads 10. The non-linear term γ ranges in $\gamma \in \{1, 2, 3\}$, likewise $p = 70\%$.

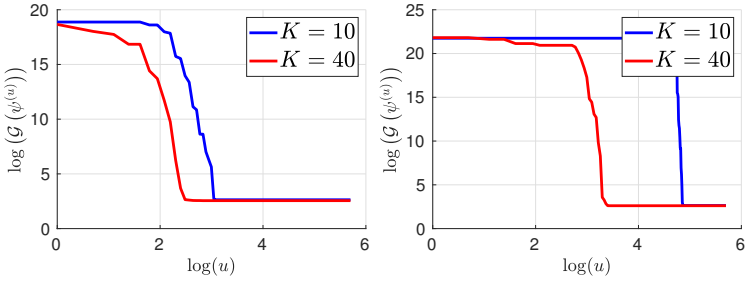
	Tabu Search Based Methods				Simulated Annealing Based Methods			
γ	U	K	TS-MGA	TS-SGA	ρ	K	SA-MGA	SA-SGA
5	100	10	0.035	0.015	0.85	10	0.037	0.018
		20	0.038			20	0.049	
		30	0.085			30	0.075	
		40	0.168			40	0.095	
	200	10	0.058		0.90	10	0.038	
		20	0.123			20	0.054	
		30	0.176			30	0.077	
		40	0.476			40	0.105	
	300	10	0.059		0.95	10	0.044	
		20	0.111			20	0.080	
		30	0.154			30	0.084	
		40	0.426			40	0.109	
6	100	10	0.018	0.014	0.85	10	0.036	0.017
		20	0.024			20	0.054	
		30	0.063			30	0.074	
		40	0.125			40	0.099	
	200	10	0.030		0.90	10	0.036	
		20	0.030			20	0.060	
		30	0.068			30	0.078	
		40	0.179			40	0.102	
	300	10	0.041		0.95	10	0.045	
		20	0.040			20	0.070	
		30	0.105			30	0.085	
		40	0.254			40	0.111	
7	100	10	0.004	0.014	0.85	10	0.042	0.017
		20	0.008			20	0.065	
		30	0.017			30	0.091	
		40	0.102			40	0.112	
	200	10	0.012		0.90	10	0.044	
		20	0.011			20	0.067	
		30	0.031			30	0.083	
		40	0.012			40	0.113	
	300	10	0.035		0.95	10	0.048	
		20	0.015			20	0.087	
		30	0.052			30	0.098	
		40	0.126			40	0.129	

Table 6: Average of elapsed times, in seconds, for the compared methods in a single assimilation step, the number of repetition reads 10. The non-linear term γ ranges in $\gamma \in \{5, 6, 7\}$, likewise $p = 70\%$.



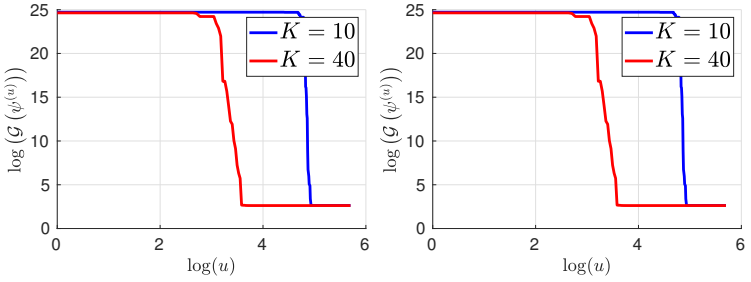
(a) $\gamma = 1$.

(b) $\gamma = 2$.



(c) $\gamma = 3$.

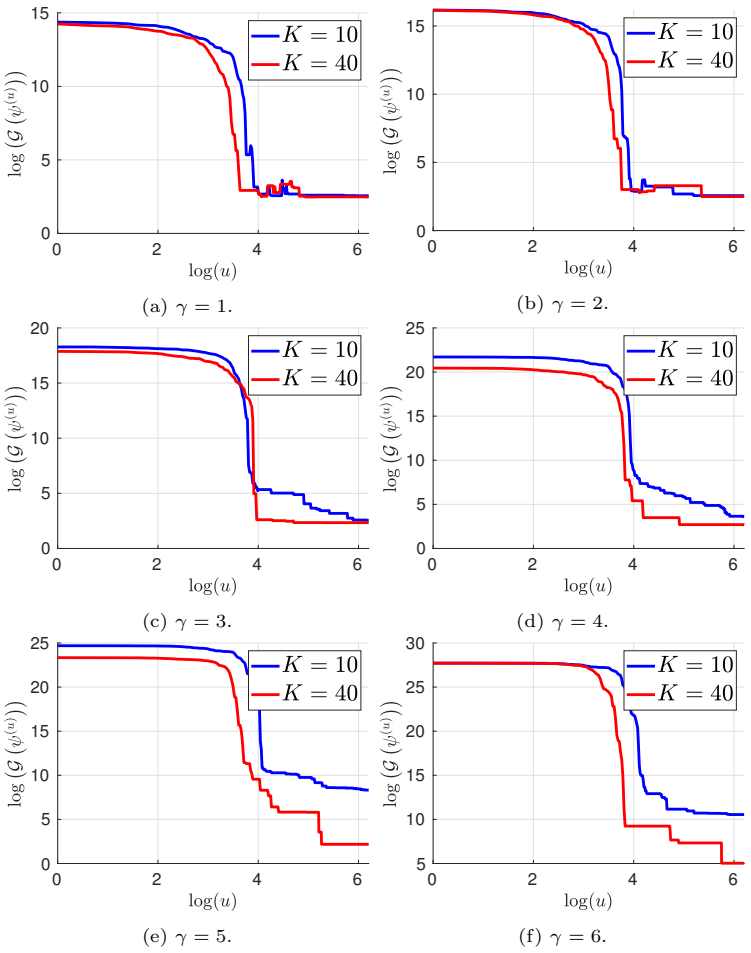
(d) $\gamma = 4$.



(e) $\gamma = 5$.

(f) $\gamma = 6$.

Figure 7: Logarithm of cost function values among iterations for a single assimilation step of the TS-MGA. Notice, as the sub-spaces dimensions are increased, the method converges faster to posterior modes of the error distribution. The number of observed components reads $p = 70\%$ and the number of iterations $U = 300$.



(a) $\gamma = 1$. (b) $\gamma = 2$.

(c) $\gamma = 3$. (d) $\gamma = 4$.

(e) $\gamma = 5$. (f) $\gamma = 6$.

Figure 8: Logarithm of cost function values among iterations for a single assimilation step of the SA-MGA method. Notice, as the sub-spaces dimensions are increased, the method converges faster to posterior modes of the error distribution (minimum values of cost functions). The number of observed components reads $p = 70\%$ while the cooling factor is set to $\rho = .95$.

5. Conclusions

Four local search methods have been proposed for the solving nonlinear data assimilation problems. The proposed methods use background states as initial seeds (solutions) of our iterative methods during assimilation steps, while observation operators are linearized about current solutions during iterations. The well-known rules in the Tabu Search and the Simulated Annealing contexts are used to update iteration formulas. Solutions are proposed, together with steepest descent approximations, for the 3D-Var cost function to reduce the number of rejected states. Sub-spaces approximations are then constructed and used in this context so as to reduce the computational effort of matrix

multiplications in full-search spaces. The global convergence of all the methods has also been theoretically proven, based on the necessary conditions and related theorems.

Experimental tests have been performed by using the standard Lorenz-96 model as our surrogate model while seven statistical models are tried to assess the accuracy and the performance of the proposed formulations. The results show that the proposed methods can reduce the prior errors by several orders of magnitudes. Even more, convergence to posterior modes can be accelerated by using sub-space approximations.

Further studies will focus on the more detailed validation of these methods using more sophisticated numerical models so as to identify if strong nonlinearity may affect the performance of the proposed approaches. In addition, it would also be useful to analyze the actual rate of convergence for different methods and to investigate how such rates of convergence may depend on the actual parameters. Furthermore, tests and validations can be carried out by using real-world data in various applications.

Acknowledgement

This work was supported in part by award UN 2018-38, and by the Applied Math and Computer Science Lab at Universidad del Norte, Colombia.

References

[1] E. D. N. Ruiz, A. Sandu, A derivative-free trust region framework for variational data assimilation, Journal of Computational and Applied Mathematics 293 (2016) 164–179 (2016).

[2] E. D. N. Ruiz, A. Sandu, J. Anderson, An efficient implementation of the ensemble kalman filter based on an iterative sherman–morrison formula, Statistics and Computing 25 (3) (2015) 561–577 (2015).

[3] S.-Y. Chang, A. Saha, Application of 3d var kalman filter in a three-dimensional subsurface contaminant transport model for a continuous pollutant source, in: Proceedings of the 2013 National Conference on Advances in Environmental Science and Technology, Springer, 2016, pp. 97–104 (2016).

[4] E. D. Nino-Ruiz, A. Sandu, X. Deng, A parallel implementation of the ensemble kalman filter based on modified cholesky decomposition, Journal of Computational Science (2017).

[5] E. D. Nino-Ruiz, A. Sandu, X. Deng, A parallel ensemble kalman filter implementation based on modified cholesky decomposition, in: Proceedings of the 6th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems, ACM, 2015, p. 4 (2015).

[6] E. D. Nino-Ruiz, A. Sandu, Efficient parallel implementation of dddas inference using an ensemble kalman filter with shrinkage covariance matrix estimation, Cluster Computing (2017) 1–11 (2017).

[7] M. Zupanski, I. M. Navon, D. Zupanski, The maximum likelihood ensemble filter as a non-differentiable minimization algorithm, Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography 134 (633) (2008) 1039–1050 (2008).

[8] A. Carrassi, S. Vannitsem, D. Zupanski, M. Zupanski, The maximum likelihood ensemble filter performances in chaotic systems, Tellus A: Dynamic Meteorology and Oceanography 61 (5) (2008) 587–600 (2008).

[9] P. Fearnhead, H. R. Künsch, Particle filters and data assimilation, Annual Review of Statistics and Its Application 5 (2018) 421–449 (2018).

[10] M. Imani, U. M. Braga-Neto, Particle filters for partially-observed boolean dynamical systems, Automatica 87 (2018) 238–250 (2018).

[11] M. Zhu, P. J. Van Leeuwen, W. Zhang, Estimating model error covariances using particle filters, Quarterly Journal of the Royal Meteorological Society 144 (713) (2018) 1310–1320 (2018).

[12] G. Evensen, The ensemble kalman filter: Theoretical formulation and practical implementation, Ocean dynamics 53 (4) (2003) 343–367 (2003).

[13] G. Evensen, Data assimilation: the ensemble Kalman filter, Springer Science & Business Media, 2009 (2009).

[14] G. Burgers, P. Jan van Leeuwen, G. Evensen, Analysis scheme in the ensemble kalman filter, Monthly weather review 126 (6) (1998) 1719–1724 (1998).

[15] P. L. Houtekamer, H. L. Mitchell, Data assimilation using an ensemble kalman filter technique, Monthly Weather Review 126 (3) (1998) 796–811 (1998).

[16] P. L. Houtekamer, H. L. Mitchell, Ensemble kalman filtering, Quarterly Journal of the Royal Meteorological Society 131 (613) (2005) 3269–3289 (2005).

[17] M. L. Stein, Limitations on low rank approximations for covariance matrices of spatial data, Spatial Statistics 8 (2014) 1–19 (2014).

[18] P. J. Bickel, E. Levina, Covariance regularization by thresholding, The Annals of Statistics (2008) 2577–2604 (2008).

[19] E. D. Nino-Ruiz, A. Mancilla, J. C. Calabria, A posterior ensemble kalman filter based on a modified cholesky decomposition (2017).

[20] E. D. Nino-Ruiz, A matrix-free posterior ensemble kalman filter implementation based on a modified cholesky decomposition, Atmosphere 8 (7) (2017) 125 (2017).

[21] J. L. Steward, J. E. Roman, A. L. Daviña, A. Aksoy, Parallel direct solution of the covariance-localized ensemble square-root kalman filter equations with matrix functions, Monthly Weather Review (2018) (2018).

[22] S. Dubinkina, Relevance of conservative numerical schemes for an ensemble kalman filter, Quarterly Journal of the Royal Meteorological Society 144 (711) (2018) 468–477 (2018).

[23] W. R. Gilks, Markov chain monte carlo, Encyclopedia of Biostatistics (2005).

[24] D. Gamerman, H. F. Lopes, Markov chain Monte Carlo: stochastic simulation for Bayesian inference, Chapman and Hall/CRC, 2006 (2006).

[25] C. M. Carlo, Markov chain monte carlo and gibbs sampling, Lecture notes for EEB 581 (2004).

[26] M. K. Cowles, B. P. Carlin, Markov chain monte carlo convergence diagnostics: a comparative review, Journal of the American Statistical Association 91 (434) (1996) 883–904 (1996).

[27] S. L. Cotter, G. O. Roberts, A. M. Stuart, D. White, Mcmc methods for functions: modifying old algorithms to make them faster, Statistical Science (2013) 424–446 (2013).

[28] A. Attia, V. Rao, A. Sandu, A hybrid monte-carlo sampling smoother for four-dimensional data assimilation, International Journal for Numerical Methods in Fluids 83 (1) (2017) 90–112 (2017).

[29] A. Attia, R. Ştefănescu, A. Sandu, The reduced-order hybrid monte carlo sampling smoother, International Journal for Numerical Methods in Fluids 83 (1) (2017) 28–51 (2017).

[30] E. D. Nino-Ruiz, C. Ardila, R. Capacho, Local search methods for the solution of implicit inverse problems, Soft Computing (2017) 1–14 (2017).

[31] E. D. Nino-Ruiz, L. E. Morales-Retat, A tabu search implementation for adaptive localization in ensemble-based methods, Soft Computing (2018) 1–17 (2018).

[32] F. Glover, Tabu searchpart ii, ORSA Journal on computing 2 (1) (1990) 4–32 (1990).

[33] F. Glover, M. Laguna, Tabu search, John Wiley & Sons, Inc., 1993 (1993).

[34] F. Glover, M. Laguna, Tabu search, in: Handbook of combinatorial optimization, Springer, 2013, pp. 3261–3362 (2013).

[35] E. Aarts, J. Korst, Simulated annealing and boltzmann machines (1988).

[36] P. J. Van Laarhoven, E. H. Aarts, Simulated annealing, in: Simulated annealing: Theory and applications, Springer, 1987, pp. 7–15 (1987).

[37] S. Kirkpatrick, Optimization by simulated annealing: Quantitative studies, Journal of statistical physics 34 (5-6) (1984) 975–986 (1984).

[38] X.-S. Yang, S. Deb, Cuckoo search via lévy flights, in: Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on, IEEE, 2009, pp. 210–214 (2009).

[39] X.-S. Yang, A new metaheuristic bat-inspired algorithm, in: Nature inspired cooperative strategies for optimization (NICSO 2010), Springer, 2010, pp. 65–74 (2010).

[40] R. D. Al-Dabbagh, F. Neri, N. Idris, M. S. Baba, Algorithmic design issues in adaptive differential evolution schemes: Review and taxonomy, Swarm and Evolutionary Computation (2018).

[41] X.-S. Yang, Nature-inspired metaheuristic algorithms, Luniver press, 2010 (2010).

[42] H. Zang, S. Zhang, K. Hapeshi, A review of nature-inspired algorithms, Journal of Bionic Engineering 7 (2010) S232–S237 (2010).

[43] A. Sotoudeh-Anvari, A. Hafezalkotob, A bibliography of metaheuristics-review from 2009 to 2015, International Journal of Knowledge-based and Intelligent Engineering Systems 22 (1) (2018) 83–95 (2018).

[44] G. N. Vanderplaats, Numerical optimization techniques for engineering design: with applications, Vol. 1, McGraw-Hill New York, 1984 (1984).

[45] S. Wright, J. Nocedal, Numerical optimization, Springer Science 35 (67-68) (1999) 7 (1999).

[46] G. Savard, J. Gauvin, The steepest descent direction for the nonlinear bilevel programming problem, Operations Research Letters 15 (5) (1994) 265–272 (1994).

[47] W. W. Hager, H. Zhang, A survey of nonlinear conjugate gradient methods, Pacific journal of Optimization 2 (1) (2006) 35–58 (2006).

[48] R. Fletcher, C. M. Reeves, Function minimization by conjugate gradients, The computer journal 7 (2) (1964) 149–154 (1964).

[49] R. M. Lewis, V. Torczon, M. W. Trosset, Direct search methods: then and now, Journal of computational and Applied Mathematics 124 (1-2) (2000) 191–207 (2000).

[50] R. Battiti, First-and second-order methods for learning: between steepest descent and newton's method, Neural computation 4 (2) (1992) 141–166 (1992).

[51] L. Grippo, F. Lampariello, S. Lucidi, A truncated newton method with nonmonotone line search for unconstrained optimization, Journal of Optimization Theory and Applications 60 (3) (1989) 401–419 (1989).

[52] V. Y. Pan, S. Branham, R. E. Rosholt, A.-L. Zheng, Newton's iteration for structured matrices, in: Fast reliable algorithms for matrices with structure, SIAM, 1999, pp. 189–210 (1999).

[53] D. F. Shanno, Conditioning of quasi-newton methods for function minimization, Mathematics of computation 24 (111) (1970) 647–656 (1970).

[54] J. Nocedal, Updating quasi-newton matrices with limited storage, Mathematics of computation 35 (151) (1980) 773–782 (1980).

[55] M. H. Loke, R. Barker, Rapid least-squares inversion of apparent resistivity pseudosections by a quasi-newton method, Geophysical prospecting 44 (1) (1996) 131–152 (1996).

[56] D. A. Knoll, D. E. Keyes, Jacobian-free newton–krylov methods: a survey of approaches and applications, Journal of Computational Physics 193 (2) (2004) 357–397 (2004).

[57] A. M. Cervantes, A. Wächter, R. H. Tütüncü, L. T. Biegler, A reduced space interior point strategy for optimization of differential algebraic systems, Computers & Chemical Engineering 24 (1) (2000) 39–51 (2000).

[58] T. G. Epperly, E. N. Pistikopoulos, A reduced space branch and bound algorithm for global optimization, Journal of Global Optimization 11 (3) (1997) 287–311 (1997).

[59] J. S. Logsdon, L. T. Biegler, A relaxed reduced space sqp strategy for dynamic optimization problems, Computers & chemical engineering 17 (4) (1993) 367–372 (1993).

[60] L. Grippo, F. Lampariello, S. Lucidi, A nonmonotone line search technique for newtons method, SIAM Journal on Numerical Analysis 23 (4) (1986) 707–716 (1986).

[61] A. Uschmajew, B. Vandereycken, Line-search methods and rank increase on low-rank matrix varieties, in: Proceedings of the 2014 International Symposium on Nonlinear Theory and its Applications (NOLTA2014), 2014, pp. 52–55 (2014).

[62] S. Hosseini, W. Huang, R. Yousefpour, Line search algorithms for locally lipschitz functions on riemannian manifolds, SIAM Journal on Optimization 28 (1) (2018) 596–619 (2018).

[63] Z.-J. Shi, Convergence of line search methods for unconstrained optimization, Applied Mathematics and Computation 157 (2) (2004) 393–405 (2004).

[64] W. Zhou, I. Akrotirianakis, S. Yektamaram, J. Griffin, A matrix-free line-search algorithm for nonconvex optimization, Optimization Methods and Software (2017) 1–24 (2017).

[65] M. Asch, M. Bocquet, M. Nodet, Data assimilation: methods, algorithms, and applications, Vol. 11, SIAM, 2016 (2016).

[66] A. Karimi, M. R. Paul, Extensive chaos in the lorenz-96 model, Chaos: An Interdisciplinary Journal of Nonlinear Science 20 (4) (2010) 043105 (2010).

[67] E. N. Lorenz, Predictability: A problem partly solved, in: Proc. Seminar on predictability, Vol. 1, 1996 (1996).